EAAP: Environmental Triggers for Asthma patients in California counties

Introduction

Commonly known across most people, asthma is a chronic disease that affects the respiratory system; however, the severity of the effects may vary as it can lead to a mild or severe case. The following project aimed to look at which environmental triggers were in correlation to spikes of hospitalizations and deaths across all the counties of California. In particular, we looked at livestock concentrations, agricultural plantations, carbon emission, and PM 2.5 concentrations since these are all major contributors to the quality of air pollution if not controlled properly.

As mentioned, the main interest of this investigation is to understand how much an impact these environmental triggers have on asthma patients. With this in mind, it might provide some insights on how we can handle regulations and policies for these contributors in their respective counties.

Data

Give our main drive to answer this question, we sought out open-source datasets that contained the needed information. For livestock, we accessed tables of information from pdf documents on NASS USDA that were then transferred onto Excel sheets. In these pdf documents, the information was structured respective to the county and indicated the number of farms with the number of livestock populations per livestock for the years of 2012 and 2017. Through this platform, we collected information on Cattle, Goats, Pigs/Hogs, Sheep/Lamb, and Poultry.

Secondly, we observed a public dataset through a named system, ‘CalEnviroscreen’, this dataset was collected in its entirety, but only certain columns were addressed as it pertained to the enterprise and commercial contributions. This dataset was able to provide the concentration values of PM 2.5, toxic waste, pesticides, and traffic particles to individual zip codes. It was then altered to group the zip codes into counties as a whole for which the concentration values were averaged across the number of instances.

Lastly, as another source of data to manage, we outreach to the community as much as we could with our proposed survey. This survey was intended to reach audiences of all ages and demographics in order to get a compiled collection of qualitative information on how they perceive the air quality in their given county. A series of background questions were asked to gauge at a general demographic followed by a series of statements that were intended to interpret the perspective of the surveyed towards a range of topics such as recycling programs, traffic moderation, company fines, consumer fines, and general air quality. From the given responses, we had to address some data quality issues where we have observations of incomplete surveys or responses that disagreed with the consent form to which we simply removed those observations from our dataset as they did not provide a proper view on the topic at hand. This allowed the dataset to be centralized on responses with a mentioned county and qualitative reactions at the minimum.

Analytical Models

Given a compiled set of datasheets, we conducted some analytical visualizations on Tableau and Excel to compare the extrema values of each situation independently and see how the factors can presumably be correlated with one another. As far as every model went, the

graphical layout of California's counties was in connection to each considered factor (i.e., Cattle population per county per year, poultry population per county per year, asthma hospitalizations/death to date per county, PM 2.5 particle concentration per county, etc.). This allowed for an independent view of each scenario to primarily gain a sense of targeted areas. The plots showed where the concentrations were the largest for each case to get a sense of what environmental factors were predominant in various regions.

For the surveys, we relied on Excel to provide some analytical findings through row-column expressions. For starters, the rows were sorted in alphabetical order with the counties as the pivot column to alphabetize. This made it easier to see the compiled responses as groups per the respective counties and obtain some statistical metrics. Primarily, a focal point of interest was to look at the perspectives of major contributors per their county of residence and see if that is in conjunction with what the data presented. Although, we did use more of the qualitative responses as well that dove into potential civic engagements for both the commercial and consumer sides. It allowed us to gauge at a sense for how a given county should have regulations made that would be favorable.

Results and Interpretations

Through our data exploration and analysis, the following indications were found:

1. Los Angeles and San Bernardino county were the counties that had reported asthma-related deaths
2. Southern California counties, as well as Central valley counties, documented asthma-related hospitalizations.
3. Tulare county (or central valley area) lead with the most cattle inventory for 2012/2017
4. Merced county (or central valley area) lead with the most goat inventory for 2017
5. Stanislaus county (or central valley area) lead with the most pigs/hogs inventory for 2012/2017
6. Merced county led with the most poultry inventory for 2012/2017; however, Riverside, San Bernardino, and San Diego counties were a near runner up
7. Kern county (or central valley area) lead with the most sheep/lamb inventory for 2012/2017
8. A large concentration of pesticide particles was documented in two Northern California counties
9. PM 2.5 has a large concentration of particles mostly in Central Valley
10. The average toxic release is low however there is a red zone in Fresno and Los Angeles counties
11. Los Angeles County leads in the traffic particle concentration

From these results, a few interpretations can be made thus far. It is showcased that hospitalizations occur more than deaths and tend to be more prevalent in Southern California and Central California counties. With these two general regions in mind, Central California counties have more livestock and agricultural dependency and that is shown across the datasets and perspectives as given from our survey responses. It is known that the dust caused from plantations during harvest season as well as the pastures given to the livestock along with their natural shedding fur contributes to the PM2.5 pollutants. As far as how much of an extent

one gives over the other can be further determined by specialists in the field to verify the educated hypothesis. Switching gears to the Southern California region, it comes to no surprise that carbon emissions alongside PM 2.5 are the main drivers for asthma triggers. This is driven by the desire to meet commercial and consumer needs. Oddly enough, for Southern California counties such as Riverside, San Bernardino, and Imperial there is a small connection to poultry livestock. As one of the counties with a leading inventory of poultry, these counties do have livestock contributors to potential asthma triggers and possibly in connection with weather patterns – not officially explored in this report.

Limitations

Our main limitation lied in our survey dataset as it was limited in responses but also prone to biased data. It was with a great attempt to push out the survey to everyone and anyone of all ages to get a broader set of perspectives; however, it tended to lie around the same age group or college students. Our main demographic was students who attend UC Merced but live back at home in different counties which is why we were able to get the perspectives in that sense. I find that this made our analysis limited since further assumptions cannot be made due to the main point of having a biased demographic providing responses. Nonetheless, there is an associative correlation as to how people feel about their county's air quality, and that definitely acted as supplemental qualitative information that was on a ranking basis.

In addition to biased data, there was also a limitation on recent data. Our open-source data was a bit outdated by about 5 years which can be considered unsupportive to the analysis. Considering it as a trend based analysis would provide a more cohesive argument as we can see over the course of years if the numbers decreased significantly and see if that was reported in the same way for the hospitalizations and deaths; however, to do so, the deaths and hospitalizations would need to be separated by year per county to see whether or not the cases for asthma severity are increasing or decreasing in connection to the documented environmental factors. In other words, a proper analytics project would ideally have the same information across the same time periods (i.e. carbon emissions per year, livestock inventory per year, hospitalizations per year); however, data is very difficult to have that clean as there will be hiccups and these at the type of data quality issues that pose as limitations for accurate finding.

Recommendations

As mentioned earlier, the survey responses have an associative correlation to the collected data as it still pertains to their perspectives on main contributors and call to actions for proper air quality treatment. First and foremost, the Central Valley counties would be having a focus on altering their livestock and agriculture regulations whereas Southern California counties would have slight regulation in the livestock/agriculture programs but have more of an emphasis on controlling consumer and commercial carbon emissions. Across our limited pool of participants, the general consensus for a starting point was that pollution was indeed a problem; however, the call to action for this issue is still mixed in perspectives. Three main points of agreements that we would also recommend as a proper proposition for asthma patients is to have financial responsibilities placed on the consumer and enterprise side. Companies should cooperate with environmental regulations such as placing a threshold on their production

otherwise a fine can be placed on them. On the other hand, it can also be placed on the consumers to pay for the change of cleaner services as companies push for the movement and financially help to some extent and more over the course of the years. There was also a strong opinion in favor of promoting more recycling programs across the counties and we would like to suggest enforcing proper etiquette with public health moral. This deals with educating the public but also keeping a rhythm to it so that it does not seem so one-sided and not solely dependent on the general public to make changes. On a last note, it was interesting to see the mixed feelings on traffic regulations, but this is yet another recommendation we feel will target the areas dealing with large amounts of carbon emission such as Los Angeles County. As it is known, LA county is densely populated and the amount of traffic and cars running on gasoline and producing emissions is extensive. This will definitely not be an overnight recommendation but more of an ongoing plan that will give more comfort and security if the federal or local government is able to take responsibility in such tasks. It is a great indication of initiative, so another potential option is to have non-profit or for-profit organizations take charge so long as it is kept consistent.

Summary

As a team we had countless group meetings to discuss, plan and work on completing every milestone leading up to the final completion of the EAAP project. From our discussions we created a project proposal, a data collection plan, a data governance plan, and conducted a survey. Having these set plans gave us a structure to follow and also enabled us to follow certain data confidentiality guidelines and procedures when handling datasets. From working on this data analytics project we gained extra knowledge on the data cleaning process, management of data governance protocols, and confidential data management. We learned that there are many other functionalities when it comes to using cloud storage services and keeping our data confidential. Furthermore, a big takeaway was learning how to deal with data quality issues within our open source datasets and our survey data.

Data Discussion

The data that we used were different datasets that reported asthma deaths and asthma hospitalization rates for the different counties across California as well as our own survey data. We also used data from the CalEnviroScreen which is a tool meant to identify California communities that are most at risk for pollution and the ones mostly being affected by it. We also looked at different agricultural data such as the amount of livestock farms and the amount of livestock sold per county. This data that we collected from public datasets was placed onto Excel sheets in order to much more easily analyze it. These sheets were ordered alphabetically by county name.

The survey we created was distributed through us, we all asked and shared the survey with people. Our target audience for the survey were people who were affected by asthma in California. Through this method we managed to obtain 36 responses. From these 36 responses, 5 of them weren’t usable. It wasn’t difficult to get people to take the survey, getting some of them to complete it was the issue.

The data was transferred from Qualtrics to an excel spreadsheet which we created with the appropriate headings that corresponded with the different questions as well as the response

IDs. The data was cleansed by seeing which responses had a low progress percentage of questions answered. If the percentage wasn't satisfactory it wasn't considered as part of the dataset. Cleansing the data did not prove to be difficult, we only needed to be cautious when reviewing the responses.

Data Architecture

For all storage of data assets we decided to use a cloud storage service, the Box which is hosted by UC Merced. We created a new file folder on the box and named it "MIST-131 Project Files." Carlos was the owner of this folder and he shared this file with every member and gave us each co-owner access. For better organization of our data assets we also created separate folders for each corresponding data asset. Such as "data" for all finalized datasets and "Raw data form" for all raw data assets etc. Every member was involved directly with the data cleaning process where we had to each convert a data asset from pdf format into excel spreadsheet. Each member was assigned an open source data set to convert and we also peer reviewed each other's work to ensure that all numbers were inputted correctly as well as identifying missing data and replacing it based on our general standard. Ricky was in charge of doing the main analysis for our data collected. To manage the versioning of data sets we had created a data documentation sheet where we would list the date of the update, file name of the dataset, the name of the member who altered the dataset, and a brief description on the changes made. In addition, on the box we have direct access to all version history and are able to revert back to prior versions at any point in time. After every new upload of the same file or alteration through the online web version of microsoft platforms, the box will automatically create a new version and list it with a version number. For instance, if it was our third upload or update the file would come up with "V3" next to it. In terms of backup procedures, we agreed that every member would store a copy of each data asset which include the documents on their personal flash drive which would be purged after completion of this data project.

Metadata Standard

For our Metadata the key elements were age, gender, ethnicity, county, and ethnicity. With age we were able to record the age group of our survey response to better understand the population. Next, gender is always important because there can be correlation between those who are affected most in a certain region by carbon emissions. Finding out what county our responses were coming from was huge to us, since our data sheets were categorized by counties. For example, when we were looking at our data sheets on farm animals, the amount of sheeps and lambs were counted based on each county in California. Lastly, ethnicity helps us again in the analytical part of our portion because we are able to find if there is a group of people who are being affected the most.

Now, our field names were all identical to the field title in our Metadata definition. We also found out what type of data type we were collecting and all of them were numbers except for "name of county". That is because we had to quantify some responses for the sake of the analytical portion of the project. For example, ethnicity can be recorded as "White", "African American" and so on, however we chose to use a number system where "1" is equivalent to "White" and "2" as "African American".

Appendix A:

- Educational Level
- Asthma Condition
- Do you know anyone with Asthma?
- Air Quality
- Causes of air pollution
- Carbon Emissions
- LiveStock Farming
- Agricultural Plantations
- Smoking
- WildFires
- Carbon Emissions (2)

The full standard of our Metadata is shown above in a list. This shows the rest of the field titles. Many of these were also quantified using the same method for “ethnicity”.

After the Project

Upon completion of the EAAP project we agreed that all data assets would remain in archive on the box. All master copies of the data that were stored on our individual flash drives would be purged. To ensure the security of the data and documents we will lock them all on the box so that only those with the password have authorized access to these files. We will also deactivate the survey to stop responses from coming in. For all data collected we will aim to anonymize all entries to the fullest potential using tactics such as hashing, permutation, noise addition, and generalization. Furthermore, we will engage in auditing procedures for reusability of the data. We agreed that the data owner, Ricky would be in charge of reviewing all reusability requests and come to a consensus within a one week period. Ricky will also enforce all members to purge the master copies of data from their flash drives and review that this has been completed via a remote standpoint. Auditing procedures for data stewards which are Aida, Carlos, and Angel will be in charge of reviewing anonymity tactics over entries and its correctness.

Reflection

After completing the assignment, we noticed some flaws that could have skewed our data. We found biases in our survey responses by not spreading out the survey to a more diverse population, especially since our research was on asthma. For example, according to our response data most of our respondents were less than 30 years old and more than 18 years old. If we had more responses from people of the third age it could be that we find more individuals with asthma related problems. Another bias is found in what counties we were surveying. Many of the counties surveyed were based on large cities where carbon emissions are more prominent compared to rural areas. A challenge we ran into was finding other datasets from different databases. As a result, most of our data sets came from the CA.gov open portal website. That is due to confidential information we could not access like, “How many people have asthma in the county of Los Angeles”. Thus, we adapted our data collection methods by finding common triggers of asthma which is why we used farm animals for data.

Data Governance Discussion

Overall, we have adhered to all the guidelines listed in our data governance plan with some extra additions. For storage purposes we did use the box as outlined above where all data assets were split into corresponding folders and put under an overall folder named “MIST-131-project files.” In addition to using the box we have also used google slides and google docs for our video presentation and project report. Both of these files were shared with every member's gmail. In terms of auditing procedures for the storage of data we have assigned the data stewards to ensure all naming conventions are following the intended format of “MIST-131_Data Assets” and “MIST-131_Data Assets_M” for all master copies as well as ensuring each asset is placed in the correct corresponding folder.

For data quality procedures as stated in the governance plan we have gone through our datasets and changed certain attributes. The gender and ethnicity entries have all changed into numbers from text and the description of the abbreviation have been added on the metadata sheet for consistency. As stated in the plan we have also ensured that the first letter of every county is capitalized. For all missing data or entries with (N/A) we have changed them into “0” for consistency and all data with a “D” was defined as disclosed and we kept those entries the same. Auditing procedures for the data quality sections were originally assigned to data stewards only but after realization that everyone was involved in the data cleaning process, every member was assigned to update all changes made on the documentation doc. Every member has followed the formatting for the documentation correctly as outlined above. We had to include the date, name of file updated, name of person who altered it, and a brief description on the changes made.

Some policies we would have wanted to add is some guidelines on what to do with empty responses that appeared in the survey data. For example, deleting policies for all missing entries within the survey data. Another policy to add would be implementing some guidelines for the documentation of metadata. It would be good to outline a format to follow as well as creating a documentation doc specifically for updated changes on the survey data. This way it allows us to easily comprehend the information when creating the metadata sheet/ data dictionary. In addition, it would be good to refine our revision and auditing policies so that data stewards have more specific rules to practice data quality concerns, updates, and auditing reports.

Ethics Discussion

Not all data collection comes without risk, especially when it involves personal questions regarding a person’s age and living area. However we lessened the risk of exposure by limiting access and also by not collecting any information that may have been too confidential. The data collected by the survey is mostly a participant’s perspective on different topics that can help or hinder air quality in their environment. There also comes another benefit in the data we collected in that there is little to no harm that can be done with its use.

Analytics projects can challenge ethical decision makers by presenting them with the choice of whether or not the data should have been collected in the first place and what is the harm that it can cause. Ethical concerns should be raised when an analytics project begins to use and share highly confidential data. Should highly confidential data be exposed to those with ill intent it could cause harm to the participants in many different aspects of their life.

Teamwork Discussion

In regards to our teamwork we kept in line with a few parts of our team agreement. As outlined in the team agreement we met during and outside of labs as necessary and also matched each other's efforts and contributions towards the project. We failed to maintain one of our agreements in which two members of the team would switch between roles after every deliverable. The only changes to team organization that would be done would have been to more properly switch the roles between the two group members who were supposed to.

Proper and good communication between team members is a very important element to any project, that especially holds true for this project. The collection of data requires proper communication in order to allow for ethical collection and proper protection of the collected data. Promotion of good communication would be a good thing to have the teams understand before starting as it will help them come to better understandings amongst each other.

Conclusion

Overall, we learned the basics of having a clear data governance plan. It becomes important to achieve consistency in terms of following structure for the guidelines we create and procedures. This makes the data governance framework tedious to work with. Since our data is always changing causing us to change data methods. For example, in our project we had to update our survey responses, which then led to updating the analytical portion of the project. Furthemore, each team member took the responsibility to review information in our data sheets as a measure to check accuracy. We like to compare data as a child because it needs constant care and maintenance in order for it to be healthy.