CSE-545 Project Report - Health and Well Being

Team - Bengaluru Big Data Boys

Rajesh Prabhakar, Rajat Hande, Ajay Gopal Krishna, Abhiram Kaushik

1. Introduction

The main goal of the project is to analyse the existing datasets and find patterns which identify the factors contributing to deaths due to **Air pollution** and **Traffic accidents** in the US and propose ideas to achieve SDG-3 Goal. Our primary focus was on premature mortality. Premature mortality is a measure of unfulfilled life expectancy. For example, in the US, the average life expectancy is 75 years and any death before the age of 75 is considered premature mortality.

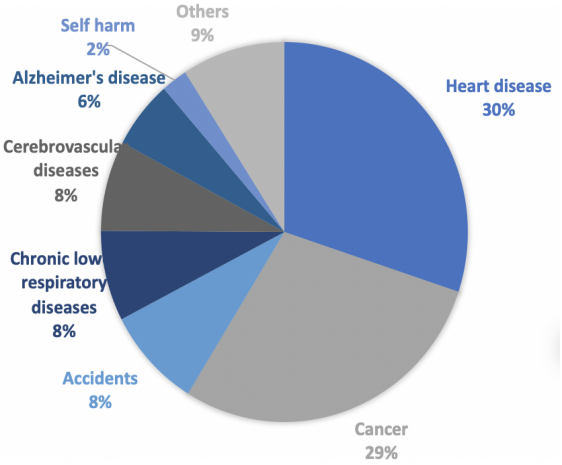

Figure 1.1 US Mortality distribution

As seen from Figure 1.1, the mortalities due to accidents and air pollution each contributes to about 8% of the total mortalities in the US, which is a significant number. The project aligns with the SDG 3 which aims to ensure healthy lives and promote well-being for all, at all ages. We have mainly focused on the following aspects of SDG:

1. By 2030, halve the number of global deaths and injuries from road traffic accidents.
2. By 2030, substantially reduce the number of deaths and illnesses from hazardous chemicals and air, water and soil pollution and contamination.

2. Background

The pollutants taken into consideration in the project are ozone (O_3), nitrogen dioxide (NO_2), carbon monoxide (CO), and sulphur dioxide(SO_2) as studies[1] associate the presence of these pollutants with the primary cause of chronic respiratory diseases. In our study, the mortality rate of chronic lower respiratory diseases is used as an indicator for deaths due to air pollution as it is evident from various studies[2], that air pollution contributes to a significant portion of the cases of such diseases. The study[3] outlines the effects of air pollution on mortality in California. We intend to extend this work to different counties in the United States and get a comprehensive perspective of the situation.

The paper[4] discusses methods to extract datasets for traffic accidents as well as the influence of environmental factors on road safety. Our study aims at providing statistical support to the various factors outlined in [5] that contribute to the road accidents in the United States and eventually help to achieve SDG 3 goals. In the report, FIPS[8] is used as a unique identifier for the counties.

3. Data

The dataset from EPA[10] provides the Air Quality Index(AQI) for different pollutants on a daily basis. The pollutant readings are taken from various sites spread across a county and need to be standardized before use. We have considered the running 8-hour averages of pollutant readings for our analysis.

IHME[11] provides datasets with estimates for age-standardized mortality rates by disease type and sex at the county level for each state. IMHE labels all the diseases with a unique number as an identifier. In our study, we have utilized chronic respiratory diseases - labelled from 508 to 520.

For preliminary analysis of factors contributing to accidents in the United States, the dataset from Kaggle[9] was used. This dataset has been sourced from MapQuest and Bing APIs[4] and contains information on accidents, weather conditions during the accidents and annotated points of interests (POIs) related to them. A comprehensive dataset related to fatal road accidents was used for our analysis. This has been obtained from the Fatality Analysis Reporting System (FARS), created in the United States by the NHTSA.

Pollution dataset: EPA[10]	1.2 GB	Daily AQI for different pollutants from 1980 - 2019
Fatal traffic accidents: FARS[12]	3.83 GB	Parameters related to Accidents, Vehicles and People involved in fatal accidents from 1970 - 2018 (1.4M records)
Accident dataset: Kaggle[9]	1 GB	Accidents data across the US: 2016-2019 (3M rows)
Chronic Respiratory Diseases Mortality Data: (GHDx[11])	700 MB	County wise mortality rates (deaths/100K) for chronic respiratory diseases from 1980-2019
Mortality (discontinued) dataset from CDC[13]	4 GB	All deaths in US from 2000-2015

Table 3.1 Outlining datasets used in the project

4. Methods

The below-described methods are performed on Google Cloud Platform. We have used a standard cluster with 1 Master (e2-standard, 2 cores, 32GB) node and 3 Worker nodes (e2-standard, 4 cores, 64GB) which primarily runs HDFS and Spark on Yarn in cluster mode.

A. Air Pollution Analysis Pipeline

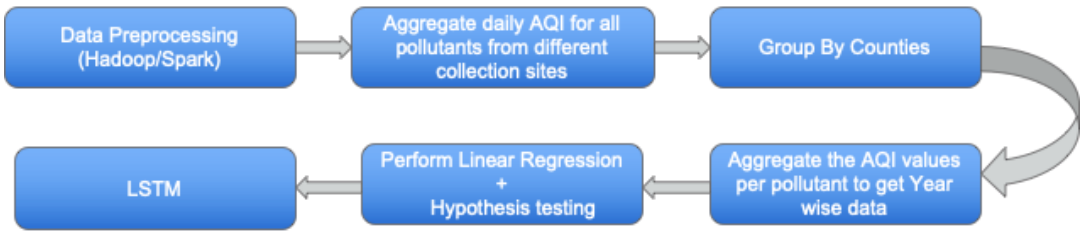

Figure 4.1.1 Air pollution analysis pipeline

The Air quality index (AQI) data from EPA was downloaded using python scripts and loaded into HDFS. Spark transformations were used to clean and process the downloaded data. In the first stage, we

loaded the AQI dataset into RDDs and filtered out the columns required for our analysis. Using Spark's filter transformation, we segregated the data based on different pollutants. In the next stage, the cleaned pollutant data was grouped by counties using Spark's reduceByKey transformations. Further, mean AQIs on a daily basis were calculated for all the counties. This process was repeated for all the pollutants and the results were saved into HDFS.

After processing the data, the daily AQI index of all pollutants for the past 40 years(1980-2019), i.e, approx. 14000 data points were obtained. We chose LSTM over other models due to the large amount of data. Parameters for the best performance are shown in Table 4.1.2.

Model details	Training	Validation	Testing	Forecast	LSTM cells	Batch Size
Parameters	11000 days	2900 days	105 days	10 days	50	10

Table 4.1.2 LSTM Parameters

The mortality dataset was cleaned using Spark's map transformations. FIPS is used as a primary key to merge the AQI and the mortality datasets through Spark's join transformation. Mean AQI per year was calculated for all the pollutants to nullify the effect of missing daily AQI values. This is possible as the AQI doesn't change drastically within a small time frame when compared to the permissible limit. As the final step, Multivariate Linear Regression and Hypothesis testing were applied on the merged dataset. The different pollutants act as feature vectors and the mortality rate is used as the dependent variable.

B. Accidents Analysis Pipeline

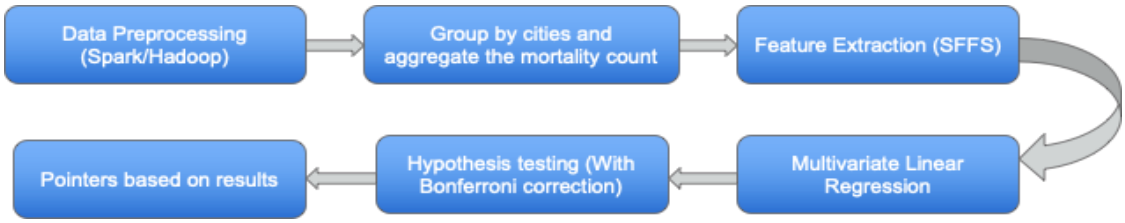

Figure 4.1.3 Accidents analysis pipeline

The accidents leading to fatalities were scraped from the FARS website and then loaded into the HDFS. Using the spark transformations the data was preprocessed to extract only features such as a number of pedestrians involved, drunk driving, hit & run etc which relate to mortality. This data was combined with the characteristics of people involved in the accident such as age group and gender. The data was grouped by cities and the daily mortalities were summed to give the total mortality count per city. For each city, the Sequential Forward Floating Selection (SFFS) algorithm was applied to extract the top 10 features influencing the fatalities. In the next stage, multivariate linear regression followed by p-value calculation for the top features selected from the SFFS stage was performed. Hypothesis testing with Bonferroni correction was applied for each of the p-values and significant features were selected. The national and the city-wise trends for the significant features were compared to analyse the impact of each feature on the mortality rate. The results for these models are explained in the next section.

5. Evaluation/Results

A. Air Pollution Analysis

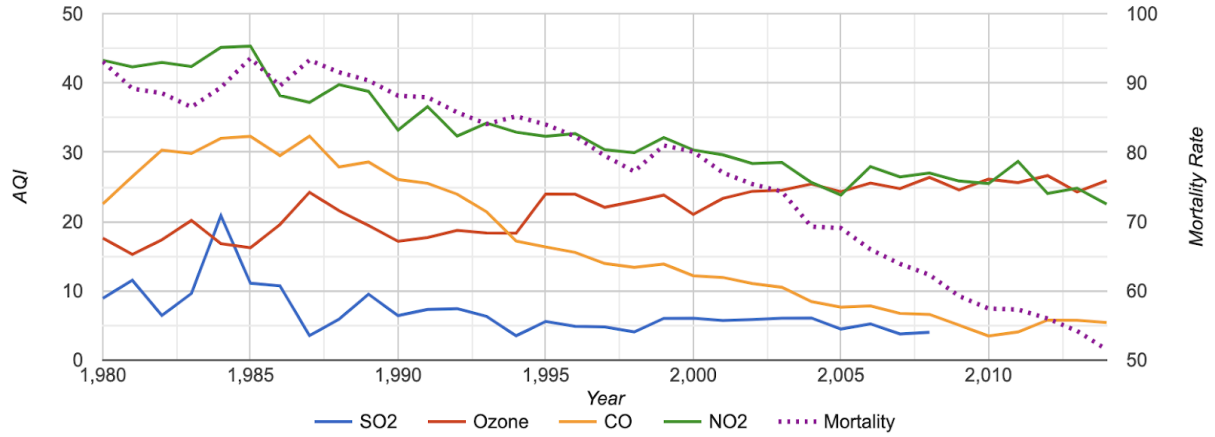

Figure 5.1.1 Air pollutants vs Mortality rates for San Francisco county (More results can be found here)

The multivariate linear regression beta values along with their Bonferroni corrected p-values are shown in Table 5.1.2. Consider the null hypothesis H_0 - "*No correlation between pollutants and deaths due to chronic diseases*". From the table 5.1.2, p-values for the pollutants O_3, NO_2, CO indicate that we can safely reject the null hypothesis. The results from our findings as shown in figure 5.1.1 are consistent with the air pollution data from San Francisco county[6]. The steady decrease in the mortality rate as seen in figure 5.1.1, can be attributed to the fact that the San Francisco Planning Department has laid out objectives and policies to help control air pollution in the county. We used LSTM to forecast the daily AQI index for the next 10 days for Los Angeles county. The model is able to predict spikes (in figure 5.1.3) which are extremely useful to find the days when the AQI suddenly crosses the permissible range.

Pollutants	β-value	p-value
O_3	-0.143	0.00022
NO_2	0.694	1.25e-18
CO	0.181	1.20e-05
SO_2	-0.104	0.0039

Table 5.1.2 p-value table

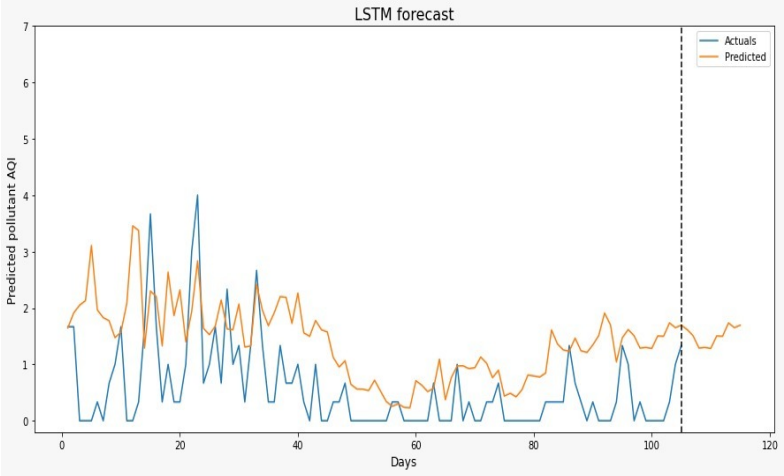

Figure 5.1.3 LSTM forecast

We can observe that the overall trends show a steady decrease in the AQI for NO_2, CO, SO_2 through the years for most of the counties in the US. However, there is a rise in the O_3 pollutant level.

From the above analyses we suggest a few pointers to reduce the mortality due to air pollution.

- The Clean Air Act must remain intact and enforced to decrease air pollution effects.
- The EPA[7] has categorized regions as "attainment" and "nonattainment" based on ozone pollutant level. In order to control Ozone pollutants EPA must closely work with states and enforce strict rules to reduce ozone pollutants.
- Using the LSTM forecast, we can send out a warning to the people to lower their activities which cause pollution. Forecasting period can also be extended to 1 month which would give ample amount of time for people to take precautionary measures.

B. Traffic Accident Analysis

The table 5.1.4 shows the Bonferroni corrected p-values for each of the top features for Los Angeles obtained after the **SFFS** feature extraction. The figure 5.1.5 shows the variation of factors that affect accident related mortality across the different cities and the entire US.

Feature	DAY_WEEK	ROUTE	A_INTER	A_INTSEC	A_ROADFC	A_JUNC	A_RD	A_HR	A_MC	A_ROLL
p-Value	2.31E-08	4.88E-1	0.349	5.07E-6	1.98E-1	1.91E-2	0.016	1.12E-2	1.15E-2	1.87E-7

Table 5.1.4 Hypothesis Testing result for Los Angeles City [*not significant*]

The figure 5.1.5 shows that the percentage of fatalities at night is more in the cities as compared to the overall US percentage. Similarly, the pedestrian involved fatalities and fatalities due to Hit and Run in the cities are almost twice and four times respectively, when compared to the overall US percentage.

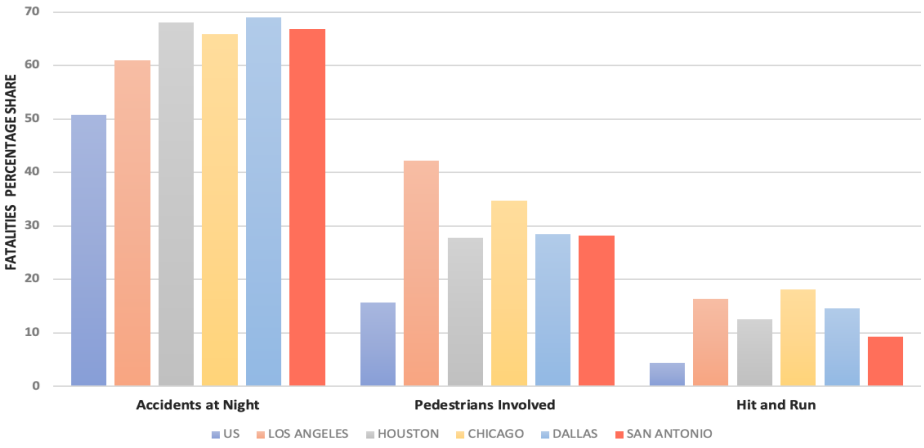

Figure 5.1.5 Fatalities due to various reasons

From the above analysis, we suggest a few pointers to curb fatalities caused by accidents.

- Promotion of Good Samaritan Law to immediately provide medical care for the accident victims as response time is a key factor in reducing accident fatalities.
- Promotion of liability insurance to reduce the Hit and Run cases.
- Pedestrian awareness and installation of traffic signals at all junctions to reduce pedestrian fatalities.
- Policing and penalties for drunk driving should be increased as it accounts for most of the accidents at night.

6. Conclusion

It is crucial to understand the collective impacts of multiple air pollutants on people's health in order to reduce the mortalities. Analysing and understanding this huge amount of data can offer a better perspective in the long term effects of air pollution and what could be the most efficient and effective measures to prevent it. The mortality from accidents can be reduced with cautious driving and by following the traffic rules and regulations by all and at all times.

Our analysis can be used by the different regulatory institutions to find the mortality based hotspots and pointers suggested by this project can be used to enforce methods and policies to reduce the overall premature mortalities.

7. References

[1] Long-term Exposure to Ambient Air Pollution and Change in Emphysema and Lung Function
[2] A review on human health perspective of air pollution with respect to allergies and asthma
[3] W209 Air Pollution Final Project
[4] https://arxiv.org/pdf/1906.05409.pdf
[5] Road Safety News and Information - Safe Roads USA
[6] https://generalplan.sfplanning.org/I10_Air_Quality.htm
[7] https://www.epa.gov/ozone-designations
[8] https://transition.fcc.gov/oet/info/maps/census/fips/fips.txt
[9] https://www.kaggle.com/sobhanmoosavi/us-accidents
[10] https://aqs.epa.gov/aqsweb/airdata/download_files.html
[11] United States Chronic Respiratory Disease Mortality Rates by County 1980-2014
[12]https://www.nhtsa.gov/node/97996/251
[13] NVSS - Mortality Data