Project Phase 2 - Group 10

João Lobato
fc62611@alunos.fc.ul.pt, Portugal

Nuno Correia
fc58638@alunos.fc.ul.pt, Portugal

Jesse Araújo
fc60578@alunos.fc.ul.pt, Portugal

1 Introduction

This report explores the problem of road accidents in the United States, aiming to understand how different factors might be related to their occurrence. To support this investigation, we analyze various datasets that could be connected to the causes of accidents, including road and weather conditions, pollution levels, and even NBA games that took place on the same day.

The richness and variety of the data allow for multiple lines of inquiry. For example, we examine whether adverse weather or road conditions contributed to specific accidents, assess whether emotional responses to NBA game outcomes (such as a loss by a home team) may have played a role, and explore if abnormal pollution levels were present on the day of the accidents.

In this first phase of the project, the following steps were completed:

(1) Definition of the problem [2];
(2) Identification and selection of relevant data sources [3];
(3) Characterization and profiling of each dataset [4];
(4) Definition of the integration plan to bring these datasets together in a meaningful way [5];
(5) Description of the tools and open-source libraries used throughout the project [10].

All data profiling, cleaning methods, blocking strategies, and similarity metric implementations were developed using **Python 3**.

2 Problem Definition

This project aims to uncover potential correlations between road accidents and various contextual factors by integrating multiple datasets. Rather than focusing solely on the circumstances of the accident itself, we seek to enrich the understanding of what external elements might contribute to its occurrence.

Specifically, we are interested in answering the following questions:

- Are there observable patterns between poor weather or road conditions and the frequency or severity of road accidents?
- Could high levels of air pollution on a given day be linked to a higher number of accidents, particularly in urban areas?
- Is there any correlation between the emotional impact of local NBA game outcomes and accident occurrences in the same region and timeframe?

These questions reflect the central idea of this phase: to explore relationships across domains that are not commonly analyzed together, and to evaluate whether meaningful insights can emerge from combining diverse datasets.

3 Data Sources

For this project, four datasets were sourced from Kaggle, each contributing distinct but complementary information that allows for a broader and more insightful analysis of road accident patterns in the United States. The datasets were chosen for their richness, relevance, and potential for integration, covering different contextual dimensions such as environmental conditions and social events. Below is a description of each dataset:

- **US Accidents (2016–2023):** This is the main dataset of the project and contains detailed records of road accidents in the United States. Each entry includes temporal and geographic information, as well as contextual features such as weather conditions, road types, visibility, and accident severity.
- **U.S. Pollution Data (2000 - 2023):** This dataset provides daily air quality measurements across various U.S. cities. It includes metrics such as carbon monoxide (CO), nitrogen dioxide (NO_2), ozone (O_3), and particulate matter (PM2.5). This data is used to examine potential correlations between pollution levels and accident occurrences.
- **NBA Game Elo and Carmelo Ratings:** This dataset records NBA game information such as team matchups, dates, results, and locations. It is used to explore the hypothesis that local emotional factors (e.g., reactions to a team loss) may influence accident trends in corresponding cities.
- **NBA Database:** This dataset contains metadata about NBA teams, including their abbreviations, full names, and cities. It serves as a supporting dataset to properly match team-related game data to the cities where road accidents occurred.

These datasets, when integrated, allow for the exploration of various external factors that may influence accident patterns, going beyond traditional accident analysis by incorporating environmental and social contexts.

4 Data Characterization

To characterize our data and better understand how to answer our problems, we created a table characterizing every column in all 4 datasets we used. All of these tables are included in the annex as well as the table of correspondences between the integrated model. A more in depth characterization is done in the dataset_characterization.ipynb file, where each dataset is looked at more closely. As for some a general characterization of each of the datasets used:

- US Accidents (2016–2023): This dataset contains 46 columns which describe where, when and how road conditions and air weather conditions were at the time a traffic accident was recorded in the US. The dates in this dataset span from 2016 to 2019. A characterization of each column is included in the annex 7. This dataset matters for our questions since we need to know if the number of accidents in a particular state increases after the team from that state loses.
- U.S. Pollution Data (2000 - 2023): This dataset serves as a complement to the weather conditions found on the US Accidents (2016–2023) dataset. It has 21 columns, and it

describes the air quality index measurements (AQI) made in a particular city in a state in the US. We utilize this dataset to better characterize what the conditions around traffic accidents were. A complete characterization of all columns can be found in the annex 6.
- NBA Game Elo and Carmelo Ratings: This dataset includes all the results from matches in the NBA spanning from 1946 to 2018. The rows in this dataset indicate matches as well as the result and the participating teams. This dataset is very important for our question since we want to see if these matches have an impact or not. A complete characterization of all columns can be found in the annex 5.
- NBA Database: This dataset serves as a complement for the NBA Game Elo and Carmelo Ratings dataset, it includes the abbreviations, full names and the state of the 30 NBA teams included in it. A complete characterization of all columns can be found in the annex 8.

4.1 Data Cleaning

To better understand possible issues in our datasets we created data profiles of these datasets and analyzed the results. While most of our data sources contain clean data, we've outlined procedures to clean the data regardless.

For all string type columns we standardized the strings by doing the following: we made them all upper case, removed all whitespace in them, removed all spaces and we removed special characters.

When it comes to duplicate rows we had two datasets with duplicate rows, these being the US Accidents (2016–2023) dataset and the U.S. Pollution Data (2000 - 2023) dataset. We dropped these rows since they add no relevant information to the problem at hand. We did this because rows have the exact date down to seconds and because of this it makes no sense to have exact copies of rows with the very same date.

Specific issues we found in the datasets were the following:

- The ‘City’ column in the US Accidents (2016–2023) dataset had some NaN values. To fill these values we used a geolocator python api. Using this API we plugged the ‘Longitude’ and ‘Latitude’ columns of the cities that had NaN values and we got the cities from these coordinates and filled in the missing values. If the geolocator api identified that the accident happened in no city (typically in cases where the accident occurred between cities/villages) we filled these cases with a “NOCITY” value;
- The ‘playoff’ column in the NBA Game Elo and Carmelo Ratings dataset had an inconsistent way to mark whether a game was a playoff match. Before 2016 playoff games were marked with a ’t’ signifying True, however from 2016 onwards playoff games are marked as ‘f’, ‘q’, ‘c’ and ‘s’. These denote finals, quarterfinals, conference finals and semifinals respectively. To simplify this, we filled all ‘t’, ‘f’, ‘q’, ‘c’ and ’s’ values with True and the rest with False. On a side note, we think that it might be interesting to analyze if losing a playoff match such as a final has a higher impact in the number of accidents than a regular match;
- The columns ‘state’ in the NBA Database dataset and ‘State’ in the US Accidents (2016–2023) and U.S. Pollution Data (2000 - 2023) datasets are checked for misspellings. We defined two arrays that include the correct spelling of the states/abbreviations and we used the Levenshtein similarity measure to correct any wrong spellings in these columns;
- The columns related to the air quality index AQI are checked for values out of its valid range. AQI values in the US can only be between 0 and 500, therefore we check for any values outside of this range and we fill them with the median of the column. We also filled NaN values with the column medians.

All the cleaning done is contained in the data_cleaning.ipynb file.

5 Integration Plan

5.1 Schema Integration

The purpose of this schema in the annex 10 is to enable cross-domain analysis by aligning records across common dimensions such as location and date, while maintaining clear separation of concerns through well-defined entities.

At the center of the schema lies the Location table, which stores the core geographical identifiers: City, State, Zipcode, Latitude, and Longitude. This table plays an important role in integrating the various datasets. It serves as the foundational entity that links pollution metrics, weather details, and accident records to a consistent geographical reference. By centralizing location information, we avoid redundancy and ensure data consistency across the schema.

The Pollution Metrics table originates from the pollution dataset and captures daily air quality indicators, including CO Mean, NO Mean, SO2 Mean, and O3 Mean. Each record is linked to a specific date and location via Location_ID, reflecting the environmental state of that place at a given time. The decision to isolate pollution data into its own table allows for a focused analysis of air quality trends and their potential correlations with other phenomena, such as traffic accidents or public health outcomes.

Weather information, although sourced from the same traffic accident dataset, has been separated into its own Weather Details table. This table stores meteorological attributes such as Temperature, Humidity, Wind Speed, Precipitation, Visibility, and Weather Condition. Each weather entry is associated with a Location_ID and a Date, allowing for precise temporal and spatial alignment with both pollution metrics and accident events. Isolating weather conditions enables cleaner data access and reusability in scenarios beyond traffic analysis—for example when assessing pollution effects under various atmospheric conditions.

The Accident table represents the core of the traffic accident dataset. It includes attributes such as Date, Severity, and the associated Location_ID and Road_ID. This structure supports detailed analysis of accidents over time and across regions. By linking each accident to its geographical location and the condition of the road where it occurred, the schema allows users to evaluate factors contributing to traffic incidents, including environmental and infrastructural variables.

Each accident is tied to a set of binary or categorical road condition indicators—such as Bump, Crossing, Junction, and Stop—through the Road Conditions table. This table stores static information associated with road infrastructure, indexed by Road_ID. Decoupling this information from the main accident table helps maintain normalization and simplifies the process of analyzing how road design contributes to accident frequency and severity.

On the sports domain side, the schema includes two additional tables: NBA Team and NBA Game. The NBA Team table holds team-level metadata, including City, State, and Abbreviation, while the NBA Game table records game-specific information such as Date, Season, Team1, Team2, scores, and playoff status. The relationship between teams and games is many-to-many, with each game involving two teams. By separating these entities, the schema supports performance tracking across time and enables advanced queries—for instance, comparing team performance relative to environmental conditions in their city on game days.

All entities are connected through well-defined relationships:

- The Location table maintains a one-to-many relationship with Pollution Metrics, Weather Details, and Accidents, enabling cross-dataset alignment through spatial and temporal dimensions.
- The Accident table links to Road Conditions via a one-to-one relationship on Road_ID.
- The NBA Game table links to the NBA Team table via team identifiers, allowing aggregation and comparative analysis.

This design ensures data normalization, reduces redundancy and enhances analytical flexibility. Moreover, the schema allows users to perform meaningful cross-domain analyses—for example, evaluating how local pollution or weather conditions on a given date may correlate with accident severity or the outcome of an NBA game in the same city.

5.2 Blocking

The first blocking strategy employed is the following: Year + City + State for the datasets US Accidents and Air Pollution. As for the NBA team elo dataset we defined the following startegy: Season + Team1 + Team2. We chose this blocking strategy because, for the US Accidents and Air Pollution datasets it allows us to quickly access rows from a certain year in a state's city. Making it so we need to do less comparisons to reach the number of accidents in a specific location. And as for the second strategy is allows us to quickly access matches from a certain year and from specific team matchups. One important aspect of this second strategy is that we have access to the state and city that a team represents. This is done through the NBA teams complementary dataset that we included. Therefore, we can form the same blocking strategy for all 3 datasets that will be used in the integration.

5.3 Similarity Metrics

To explore the relationships between datasets and understand potential alignment or correlation points, we applied three different similarity metrics—Jaccard similarity, Euclidean distance, and Cosine similarity—each tailored to different types of data features and comparisons.

(1) Jaccard Similarity (City Names across Datasets): The goal here was to measure textual overlap between city names in the Accidents and Pollution datasets. Using Jaccard similarity on binarized token representations of city names, we identified cities that are lexically similar or potentially represent the same location despite formatting differences. This helps in aligning and linking records from different datasets where direct city name matches are inconsistent due to variations in naming conventions.

(2) Euclidean Distance (Severity and Temperature): For numerical comparison, we computed the Euclidean distance on normalized Severity and Temperature(F) data within the Accidents dataset. This metric provides information on how similar or different incidents are in terms of intensity and environmental conditions. Smaller distances indicate clusters of similar events, which can be valuable for pattern detection or clustering analyses.

(3) Cosine Similarity (Pollution vs. Weather Profiles): Finally, we assessed the alignment between air pollution profiles (from the Pollution dataset) and corresponding weather conditions (from the Accidents dataset) using cosine similarity. This metric measures the directional similarity between two multidimensional feature vectors, allowing us to see whether pollution patterns are reflected in local weather conditions. Higher cosine similarity suggests that environmental factors might be closely associated with pollution levels in specific time frames or locations.

Each one of these similarity measures brings a unique lens to our analysis: Jaccard connects location labels, Euclidean uncovers intra-dataset structure, and Cosine explores inter-dataset feature alignment

6 Improvements to Phase 1

6.1 Blocking changes

We changed the blocking strategy used for the US Accidents dataset, instead of Year + City + State we now use Date + State. We made this change to get more direct matches; since the US Accidents dataset is too big we need to reduce the number of comparisons a lot more.

6.2 Entity Matching changes

Different similarity metrics were used for entity matching, and these new similarity metrics were also updated to be used in contexts that make more sense. The changes were the following:

(1) The Jaro similarity measure was used for entity matching between the NBA Game Elo dataset and the NBA Database dataset. It was used with team abbreviations to match teams with their respective state. Jaro was chosen for this because it is designed for small strings, and all of the abbreviations were 3 characters long.
(2) A combination of Jaro + Levenshtein was used to find matches between NBA loser teams' states and cities and the states and cities in which accidents took place. The Jaro similarity measure was used to find matches between state abbreviations, again using the same logic as aforementioned. The Levenshtein measure was also used to compare the differences between city names, since these were bigger.
(3) The Jaro similarity measure was used to match state abbreviations between the U.S. Pollution dataset and the US Accidents dataset. Jaro was selected in this context because of the reasons outlined previously, particularly its effectiveness when applied to short-string comparisons such as state abbreviations.
(4) The fuzzywuzzy similarity measure was used to look for matching City names when matching between the Accidents dataset and the Pollution dataset using a 0.90 similarity threshold. The result of these matches was saved under the City_Resolved column. Fuzzywuzzy was selected due to its efficacy when comparing strins that can contain small variations such as spaces, uppercase/lowercase or typos.

7 Data marts

We defined 3 data marts, one for each of the questions we defined in Section [2]. We provide a short description of the contents of each data mart:

(1) NBA-Accidents data mart: This data mart includes all the information from accidents that occurred in days were an NBA team lost. The information contained in it is the following: description of each accident, state and city in which it occured, the NBA team that lost along with the winner, the date and also the exact timestamp when the accident occurred. To obtain this data mart first we did entity matching between the NBA Elo Dataset and the NBA Database to obtain the state and city of each team, we used the team abbreviations along with the Jaro similarity measure. Next up we did entity matching between the NBA Game Elo dataset and the US Accidents one. To obtain matches between these two datasets we used a combination of Jaro + Levenshtein in the state abbreviations (Jaro) and the state (Levenshtein).
(2) Accidents-Pollution data mart: This data mart includes all the information from accidents that occurred on the same days and locations where pollution data was collected. The data mart contains detailed information about each accident, such as its severity, location (city and state), timestamp, and description, along with pollution measurements recorded in that city on the same date. To create this data mart, we first performed entity matching between the pollution dataset and the US Accidents dataset. In the first fusion step, we selected the most recent pollution measurement for each city-state and date, to ensure that we kept the most relevant air quality record. In the second step, we filtered the accidents to include only those that occurred in cities where pollution stations were known to exist. This allowed us to build a reliable and geographically consistent data mart for analyzing the relationship between pollution levels and accident severity.
(3) Weather-Road Conditions data mart: This data mart has all the information about traffic accidents that occurred in certain places along with the date corresponding to registered measurements of air pollution. The data mart includes detailed information about each accident including severity, location (city and state), date, timestamp. It also includes temperature, humidity and the mean values of air pollutants (NO_2, CO, O_3, SO_2) measured in the location where it happened. To construct this data mart, there was an initial fusion between the data of the traffic accidents and the air pollution dataset using the Date and City columns as merging keys. Data fusion strategies were used to solve conflicts with common attributes.

8 Data Fusion Strategies

We defined six data fusion strategies, we needed to use these after doing Entity Matching to disambiguate entities which caused conflicts, either due to referencing the same entity or just duplicates from the matching. The data fusion strategies we used were as follows:

(1) If more than two teams match with the same abbreviation, we keep the most recently founded team's information. This is because there are some old teams whose match data will not match any of the US accidents data. This happens since US Accidents has data from 2016 to 2018 and the NBA Game Elo has matches from before 2016.
(2) After merging on blocking and doing entity matching between NBA loser team's states and cities and accident states and cities. Some of these refer to the same exact accident, down the the same timestamp. Our strategy is that we keep only one row to avoid including duplicates of the same accident. This was we avoid inflating the numbers in favor of there being more accidents in NBA game days.
(3) If there were multiple pollution records for the same city and date, we kept only the most recent one. This was done to ensure that each accident would be matched with the

most up-to-date pollution data available for that location, since older pollution readings on the same day may not reflect the conditions at the time of the accident.

(4) After identifying all cities with pollution monitoring stations, we filtered the accident data to keep only those accidents that occurred in the same cities and states as those stations. This step was necessary to ensure spatial alignment between pollution observations and accidents, avoiding the inclusion of unmatched or geographically inconsistent records.
(5) After merging pollution and accident records, we identified cases where multiple rows referred to the same accident based on identical timestamps. In these situations, we kept only the first occurrence to avoid duplicate records of the same event. This step ensured the accuracy of the data mart by preventing inflated accident counts in the analysis.
(6) Slot filling was used in the Weather-Road Conditions data mart to fill missing City and State values which were present in one but not in the other.
(7) Conflict resolution was used in the Weather-Road Conditions data mart to keep the base values of the pollution measurements in cases where there were divergences in them.

9 Data Analytics

9.1 Visualizations developed

To explore our questions mentioned above, we developed various visualizations to help us do so. The following visualizations were developed:

(1) Time-Series plots to look at trends in the number and proportion of accidents in the years 2016, 2017 and 2018 per month.
(2) A map of the US that colors the states in a hot-cold gradient depending on if the difference in proportion between accidents on game losses and non-game day.
(3) A histogram with the distribution of the number of accidents per state in game day losses and non game day losses.
(4) A box-plot between accidents on game losses and non-game day.
(5) Bar plots showing the distribution of severity levels on high and low pollution days for each pollutant.
(6) Heatmap showing average pollutant levels across severity levels.
(7) Table listing the top 10 states by average severity with corresponding pollution values.

9.2 Critical Reflection About The Results

When it comes to reflecting on the different results in order to answer our 3 questions, we concluded the following:

(1) From 2016-2018 there are months that appear to have a higher proportion in accidents in NBA game day losses when compared to non-game day accidents. November and December both have higher proportion on game days in 2017 and 2016. And January, February and March as well in 2017 and 2018.
(2) When we take a loot at the difference in proportions state wise, California across all three years consistently has a higher proportion of accidents in game day losses. We believe that this is due to the fact that there are 4 teams, which is the highest number of any state, from California meaning that losses from Californian team matched with more accidents. Texas and Florida also tend to have a higher proportion of accidents in game day losses. And the rest of the states tend to have differences that are very small or they just out right have a higher proportion in non game days. It's also worth noting that Texas has 3 teams meaning it's the second state with the most team, whence why it shows a trend similar to California.
(3) The bar charts show small differences in accident severity between high and low pollution days. For (NO_2, SO_2), there's a slight increase in level 3 severity accidents on high-pollution days. (CO_2, O_3) show minor variations but follow similar patterns. These shifts may suggest pollution plays a role, but we can't conclude it's the cause, as other factors like traffic or weather might also influence accident severity.
(4) The heatmap shows (NO_2, SO_2) levels increase with severity, but (NO_2) is especially low (6.6) at severity 1 compared to 11 at higher levels. This contrast might suggest a threshold effect or differing exposure conditions. Still, other variables could explain this pattern.
(5) According to the table, states like Illinois, Mississippi, and Rhode Island show some of the highest average accident severities (above 2.66), with varying pollution levels. Connecticut, despite having the highest (NO_2) value (12.88), has a lower severity average (2.46). On the other hand, Colorado combines both high severity (2.44) and the highest (NO_2) (18.43), reinforcing what we see in the heatmap. This highlights that while patterns exist, they are not entirely consistent across states.
(6) The data shows little correlation between the severity of accidents and atmospheric measurement . When it comes the time of day of these accidents we noticed that the data showed more accidents happened in the morning which could be related to fact that there could be more cars in the morning. However, most of these accidents were of low severity.

9.3 Limitations and Other Factors

Some limitations and other factors that could be at play when it comes to the results found in our visualizations are:

(1) Results seen from the use of traffic accidents data can be affected by other social events that might've happened, influencing the number of people on the road which can lead to there being more accidents.
(2) We used Time-Series plots per month, this means that trends such as holidays could also have an affect on the number of accidents in a given month. Since if there is a big holiday in a month, more people will be on the road, making it so there can be more opportunities for there to be accidents.

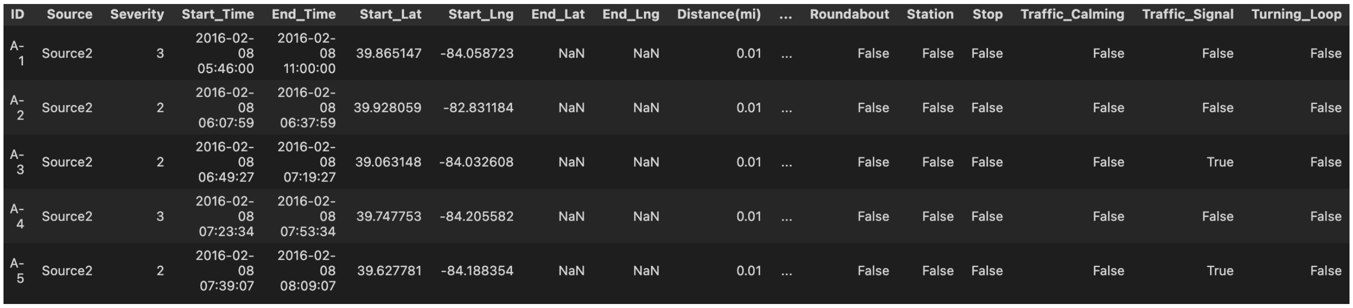
(3) Severity classification may be influenced by reporting inconsistencies or local enforcement standards, which can affect comparability across regions.

10 Tools and Libraries

The libraries and open sources used in this phase were:

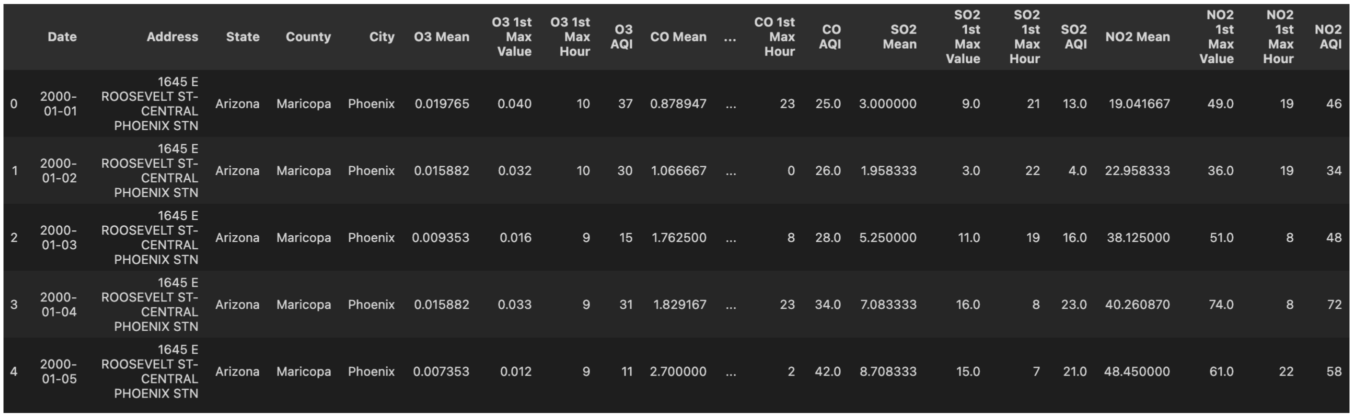
- pandas: used to read and manage the datasets;
- ydata_profiling: Used to obtain the data profiling of each of the datasets;
- re: Utilized to match strings that had special characters, so that we could remove them;
- py_stringmatching: Utilized to compute the Levenshtein and Jaro similarity measures between strings, so that we could fix misspellings in the data on certain columns;
- geopy: Utilized to obtain the cities/villages that had an NaN through their longitude and latitude values;
- CountVectorizer: Transforms text data (e.g., city names) into binary vector representations for computing similarity measures;
- jaccard_score: Computes the Jaccard similarity between two binary vectors, helping compare the overlap between tokenized strings;
- MinMaxScaler: Scales numeric features to a normalized range [0, 1], ensuring comparability across different metrics;
- pdist: Computes pairwise distances between rows in a dataset, used to measure similarity or dissimilarity between events;
- squareform: Converts the pairwise distance output from pdist into a readable matrix format;
- cosine_similarity: Calculates the cosine similarity between two sets of normalized vectors, often used to compare high-dimensional data like pollution and weather profiles.
- numpy: Used for efficient numerical operations such as array transformations and conditional assignments;
- process: From the rapidfuzz library, used to perform fuzzy matching between city or city-state strings in different datasets;
- fuzz: Also from rapidfuzz, provides similarity scorers such as Levenshtein ratio for comparing and aligning text entries;
- plot: Refers to the pyplot module from matplotlib, used to create and customize data visualizations;
- seaborn: Used to produce statistical plots such as heatmaps, histograms, boxplots, and countplots for analytical insight.
- plotly: Used to create the map with the US.

A Annex



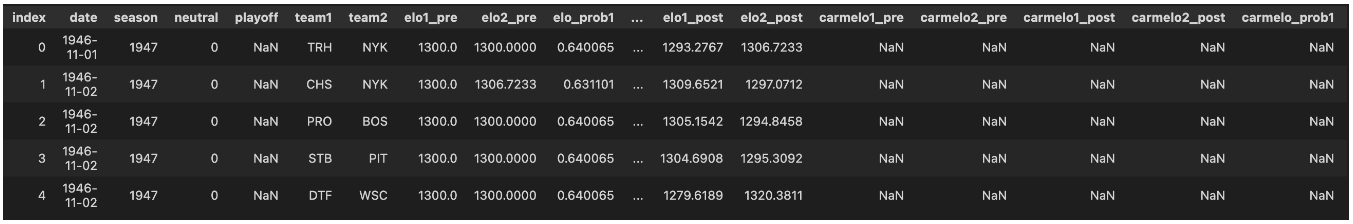
ID	Source	Severity	Start_Time	End_Time	Start_Lat	Start_Lng	End_Lat	End_Lng	Distance(mi)	...	Roundabout	Station	Stop	Traffic_Calming	Traffic_Signal	Turning_Loop
A-1	Source2	3	2016-02-08 05:46:00	2016-02-08 11:00:00	39.865147	-84.058723	NaN	NaN	0.01	...	False	False	False	False	False	False
A-2	Source2	2	2016-02-08 06:07:59	2016-02-08 06:37:59	39.928059	-82.831184	NaN	NaN	0.01	...	False	False	False	False	False	False
A-3	Source2	2	2016-02-08 06:49:27	2016-02-08 07:19:27	39.063148	-84.032608	NaN	NaN	0.01	...	False	False	False	False	True	False
A-4	Source2	3	2016-02-08 07:23:34	2016-02-08 07:53:34	39.747753	-84.205582	NaN	NaN	0.01	...	False	False	False	False	False	False
A-5	Source2	2	2016-02-08 07:39:07	2016-02-08 08:09:07	39.627781	-84.188354	NaN	NaN	0.01	...	False	False	False	False	True	False

Figure 1: Head of the US Accidents (2016–2023) dataset.



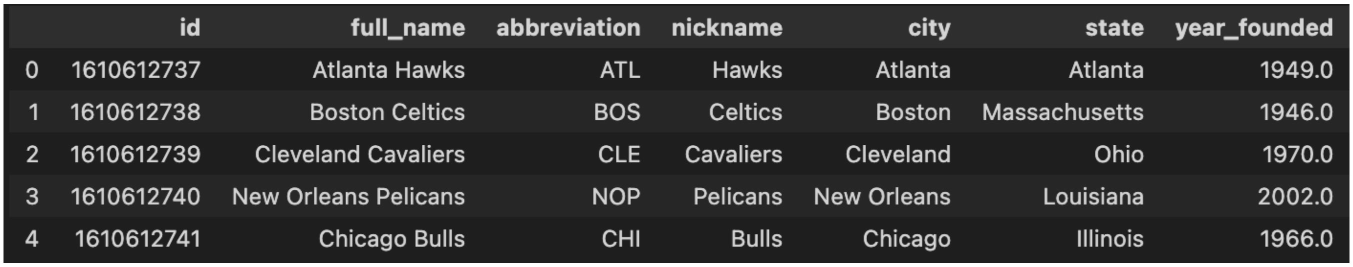
	Date	Address	State	County	City	O3 Mean	O3 1st Max Value	O3 1st Max Hour	O3 AQI	CO Mean	...	CO 1st Max Hour	CO AQI	SO2 Mean	SO2 1st Max Value	SO2 1st Max Hour	SO2 AQI	NO2 Mean	NO2 1st Max Value	NO2 1st Max Hour	NO2 AQI
0	2000-01-01	1645 E ROOSEVELT ST-CENTRAL PHOENIX STN	Arizona	Maricopa	Phoenix	0.019765	0.040	10	37	0.878947	...	23	25.0	3.000000	9.0	21	13.0	19.041667	49.0	19	46
1	2000-01-02	1645 E ROOSEVELT ST-CENTRAL PHOENIX STN	Arizona	Maricopa	Phoenix	0.015882	0.032	10	30	1.066667	...	0	26.0	1.958333	3.0	22	4.0	22.958333	36.0	19	34
2	2000-01-03	1645 E ROOSEVELT ST-CENTRAL PHOENIX STN	Arizona	Maricopa	Phoenix	0.009353	0.016	9	15	1.762500	...	8	28.0	5.250000	11.0	19	16.0	38.125000	51.0	8	48
3	2000-01-04	1645 E ROOSEVELT ST-CENTRAL PHOENIX STN	Arizona	Maricopa	Phoenix	0.015882	0.033	9	31	1.829167	...	23	34.0	7.083333	16.0	8	23.0	40.260870	74.0	8	72
4	2000-01-05	1645 E ROOSEVELT ST-CENTRAL PHOENIX STN	Arizona	Maricopa	Phoenix	0.007353	0.012	9	11	2.700000	...	2	42.0	8.708333	15.0	7	21.0	48.450000	61.0	22	58

Figure 2: Head of the U.S. Pollution Data (2000 - 2023) dataset.



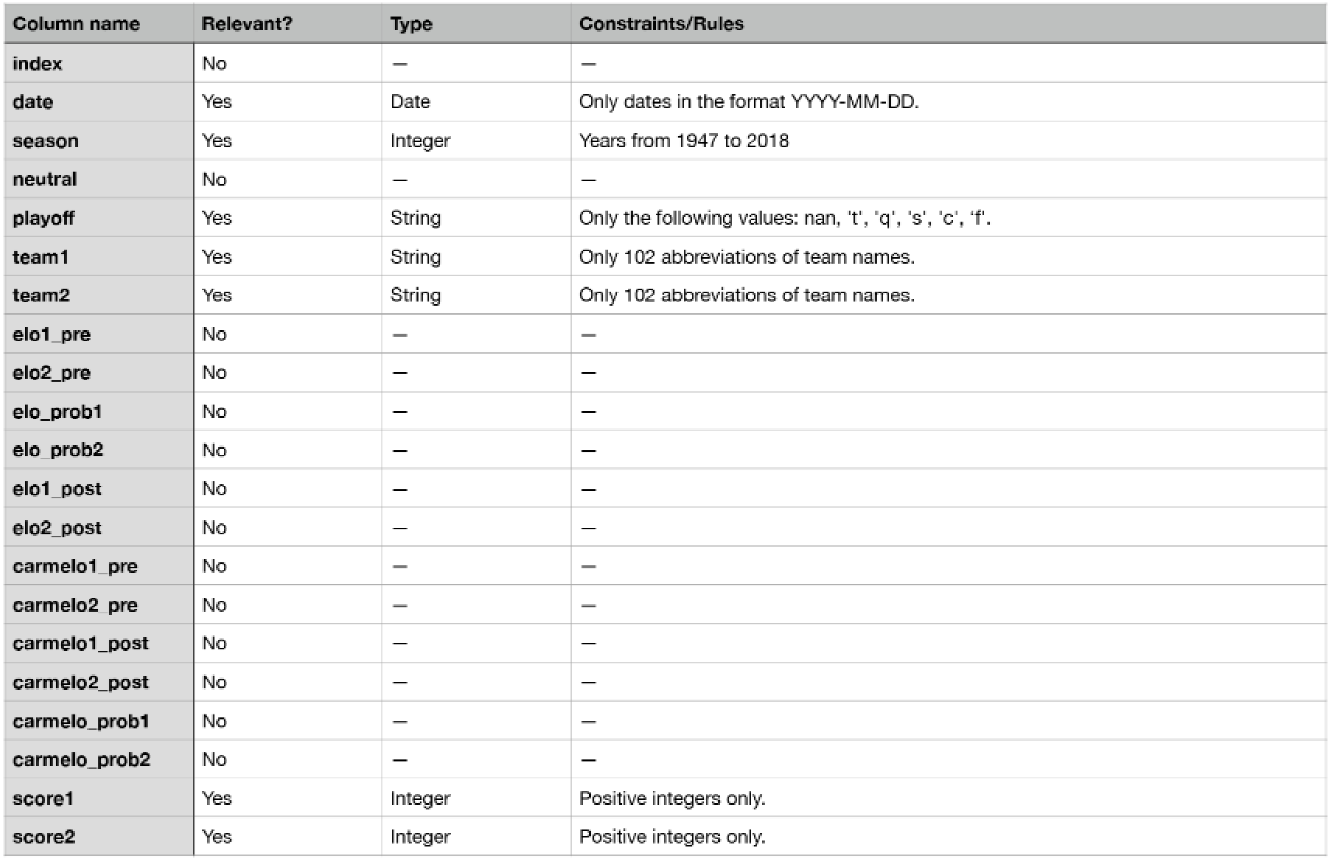
index	date	season	neutral	playoff	team1	team2	elo1_pre	elo2_pre	elo_prob1	...	elo1_post	elo2_post	carmelo1_pre	carmelo2_pre	carmelo1_post	carmelo2_post	carmelo_prob1
0	1946-11-01	1947	0	NaN	TRH	NYK	1300.0	1300.0000	0.640065	...	1293.2767	1306.7233	NaN	NaN	NaN	NaN	NaN
1	1946-11-02	1947	0	NaN	CHS	NYK	1300.0	1306.7233	0.631101	...	1309.6521	1297.0712	NaN	NaN	NaN	NaN	NaN
2	1946-11-02	1947	0	NaN	PRO	BOS	1300.0	1300.0000	0.640065	...	1305.1542	1294.8458	NaN	NaN	NaN	NaN	NaN
3	1946-11-02	1947	0	NaN	STB	PIT	1300.0	1300.0000	0.640065	...	1304.6908	1295.3092	NaN	NaN	NaN	NaN	NaN
4	1946-11-02	1947	0	NaN	DTF	WSC	1300.0	1300.0000	0.640065	...	1279.6189	1320.3811	NaN	NaN	NaN	NaN	NaN

Figure 3: Head of the NBA Game Elo and Carmelo Ratings dataset.



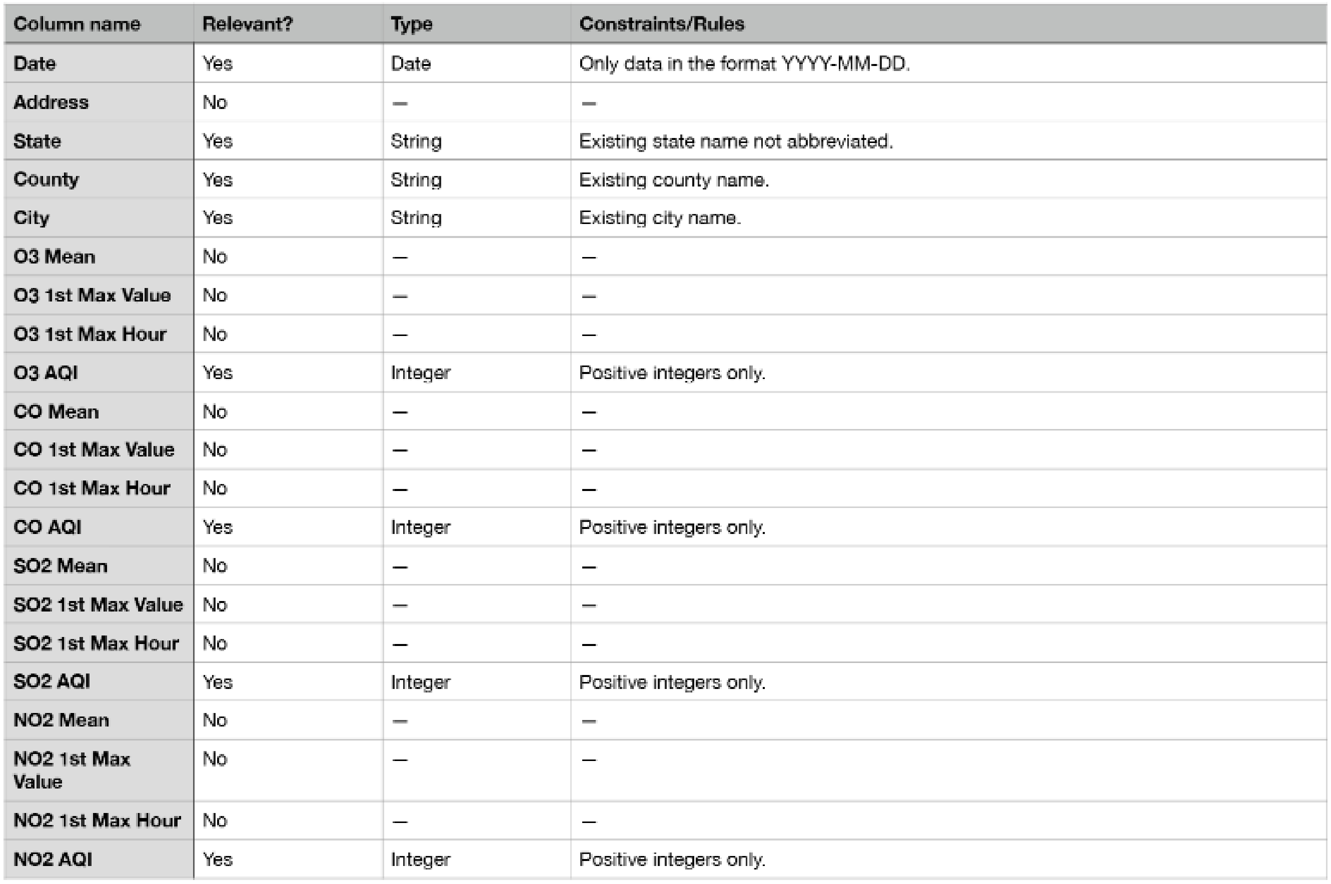
	id	full_name	abbreviation	nickname	city	state	year_founded
0	1610612737	Atlanta Hawks	ATL	Hawks	Atlanta	Atlanta	1949.0
1	1610612738	Boston Celtics	BOS	Celtics	Boston	Massachusetts	1946.0
2	1610612739	Cleveland Cavaliers	CLE	Cavaliers	Cleveland	Ohio	1970.0
3	1610612740	New Orleans Pelicans	NOP	Pelicans	New Orleans	Louisiana	2002.0
4	1610612741	Chicago Bulls	CHI	Bulls	Chicago	Illinois	1966.0

Figure 4: Head of the NBA Database dataset.



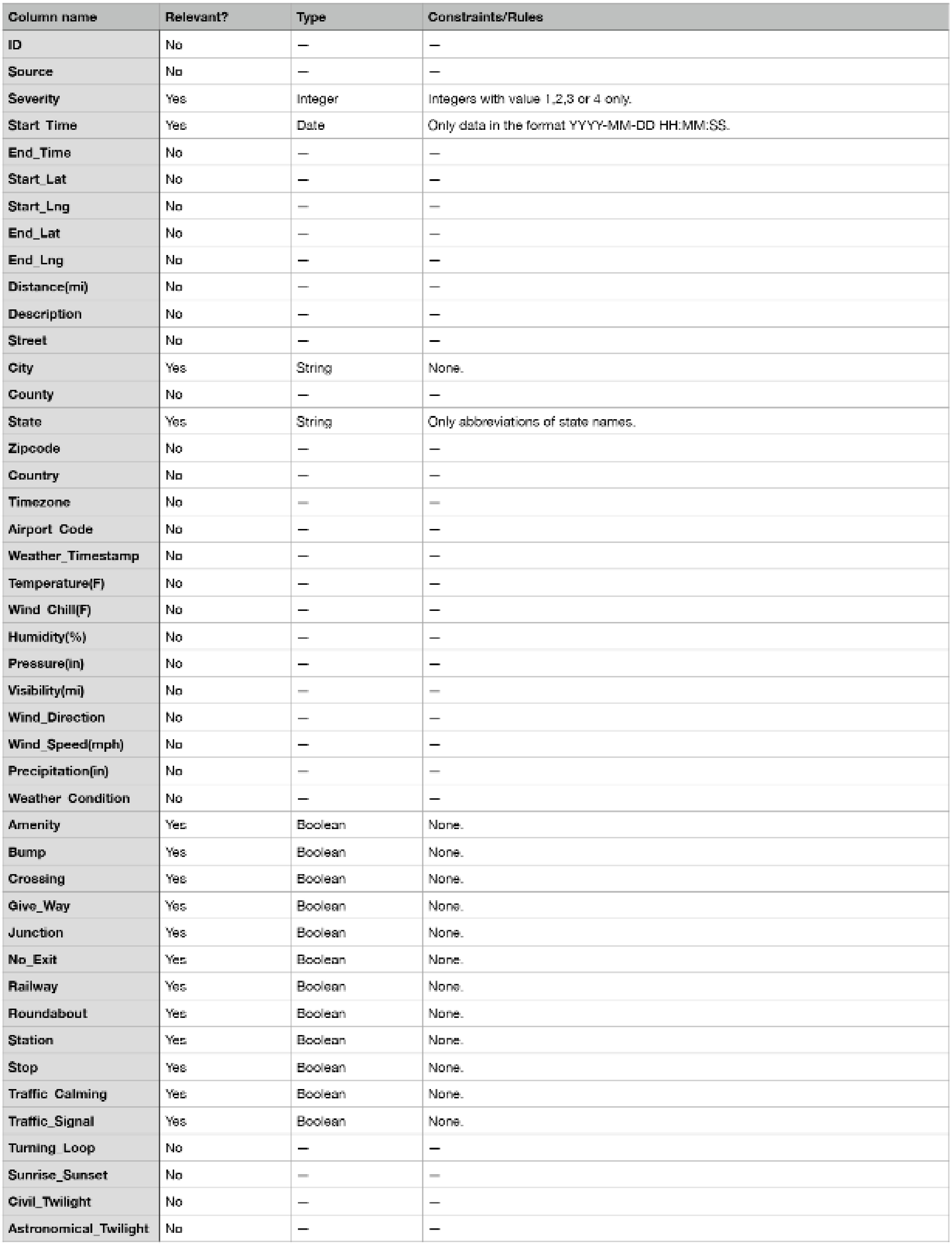
Column name	Relevant?	Type	Constraints/Rules
index	No	—	—
date	Yes	Date	Only dates in the format YYYY-MM-DD.
season	Yes	Integer	Years from 1947 to 2018
neutral	No	—	—
playoff	Yes	String	Only the following values: nan, 't', 'q', 's', 'c', 'f'.
team1	Yes	String	Only 102 abbreviations of team names.
team2	Yes	String	Only 102 abbreviations of team names.
elo1_pre	No	—	—
elo2_pre	No	—	—
elo_prob1	No	—	—
elo_prob2	No	—	—
elo1_post	No	—	—
elo2_post	No	—	—
carmelo1_pre	No	—	—
carmelo2_pre	No	—	—
carmelo1_post	No	—	—
carmelo2_post	No	—	—
carmelo_prob1	No	—	—
carmelo_prob2	No	—	—
score1	Yes	Integer	Positive integers only.
score2	Yes	Integer	Positive integers only.

Figure 5: Column characterization in the NBA Game Elo and Carmelo Ratings dataset.



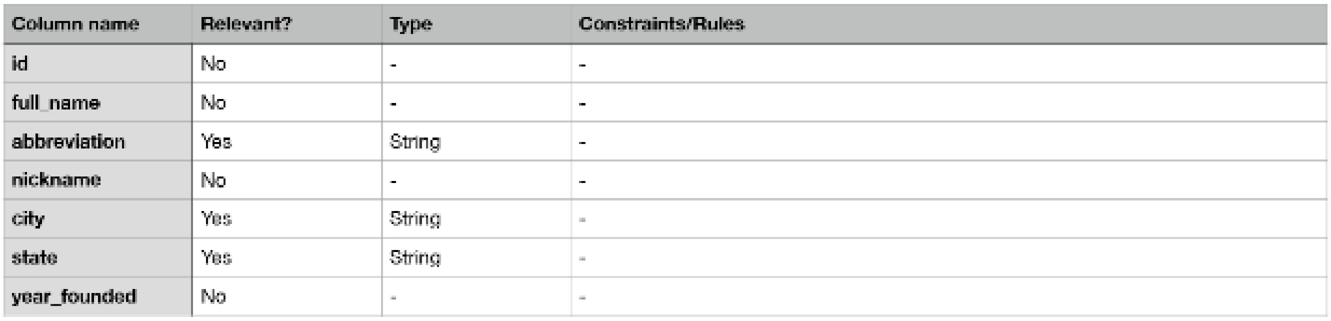
Column name	Relevant?	Type	Constraints/Rules
Date	Yes	Date	Only data in the format YYYY-MM-DD.
Address	No	—	—
State	Yes	String	Existing state name not abbreviated.
County	Yes	String	Existing county name.
City	Yes	String	Existing city name.
O3 Mean	No	—	—
O3 1st Max Value	No	—	—
O3 1st Max Hour	No	—	—
O3 AQI	Yes	Integer	Positive integers only.
CO Mean	No	—	—
CO 1st Max Value	No	—	—
CO 1st Max Hour	No	—	—
CO AQI	Yes	Integer	Positive integers only.
SO2 Mean	No	—	—
SO2 1st Max Value	No	—	—
SO2 1st Max Hour	No	—	—
SO2 AQI	Yes	Integer	Positive integers only.
NO2 Mean	No	—	—
NO2 1st Max Value	No	—	—
NO2 1st Max Hour	No	—	—
NO2 AQI	Yes	Integer	Positive integers only.

Figure 6: Column characterization in the U.S. Pollution Data (2000 - 2023) dataset.



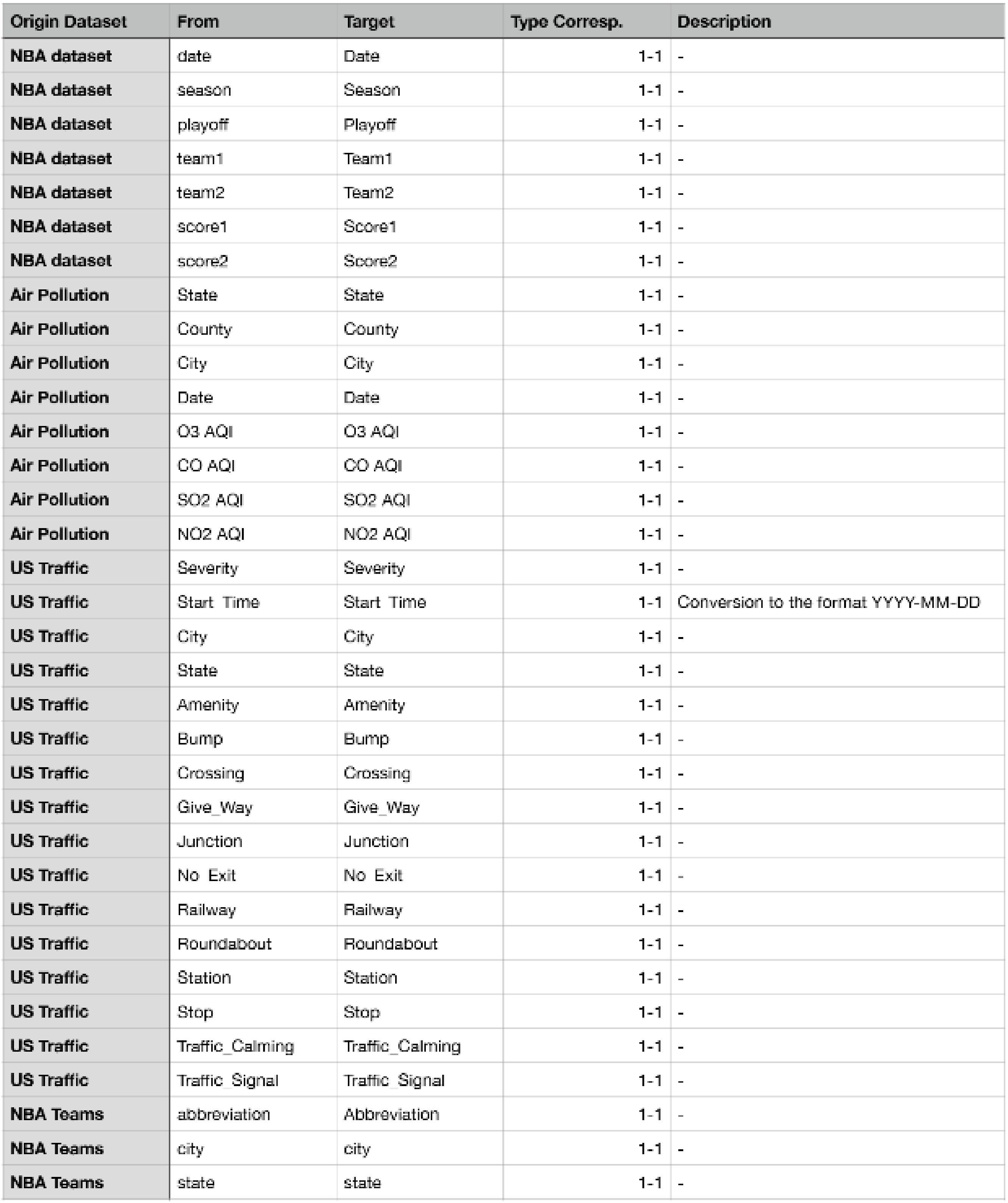
Column name	Relevant?	Type	Constraints/Rules
ID	No	—	—
Source	No	—	—
Severity	Yes	Integer	Integers with value 1,2,3 or 4 only.
Start_Time	Yes	Date	Only data in the format YYYY-MM-DD HH:MM:SS.
End_Time	No	—	—
Start_Lat	No	—	—
Start_Lng	No	—	—
End_Lat	No	—	—
End_Lng	No	—	—
Distance(mi)	No	—	—
Description	No	—	—
Street	No	—	—
City	Yes	String	None.
County	No	—	—
State	Yes	String	Only abbreviations of state names.
Zipcode	No	—	—
Country	No	—	—
Timezone	No	—	—
Airport_Code	No	—	—
Weather_Timestamp	No	—	—
Temperature(F)	No	—	—
Wind_Chill(F)	No	—	—
Humidity(%)	No	—	—
Pressure(in)	No	—	—
Visibility(mi)	No	—	—
Wind_Direction	No	—	—
Wind_Speed(mph)	No	—	—
Precipitation(in)	No	—	—
Weather_Condition	No	—	—
Amenity	Yes	Boolean	None.
Bump	Yes	Boolean	None.
Crossing	Yes	Boolean	None.
Give_Way	Yes	Boolean	None.
Junction	Yes	Boolean	None.
No_Exit	Yes	Boolean	None.
Railway	Yes	Boolean	None.
Roundabout	Yes	Boolean	None.
Station	Yes	Boolean	None.
Stop	Yes	Boolean	None.
Traffic_Calming	Yes	Boolean	None.
Traffic_Signal	Yes	Boolean	None.
Turning_Loop	No	—	—
Sunrise_Sunset	No	—	—
Civil_Twilight	No	—	—
Astronomical_Twilight	No	—	—

Figure 7: Column characterization in the US Accidents (2016–2023) dataset.



Column name	Relevant?	Type	Constraints/Rules
id	No	-	-
full_name	No	-	-
abbreviation	Yes	String	-
nickname	No	-	-
city	Yes	String	-
state	Yes	String	-
year_founded	No	-	-

Figure 8: Column characterization in the NBA Database dataset.



Origin Dataset	From	Target	Type Corresp.	Description
NBA dataset	date	Date	1-1	-
NBA dataset	season	Season	1-1	-
NBA dataset	playoff	Playoff	1-1	-
NBA dataset	team1	Team1	1-1	-
NBA dataset	team2	Team2	1-1	-
NBA dataset	score1	Score1	1-1	-
NBA dataset	score2	Score2	1-1	-
Air Pollution	State	State	1-1	-
Air Pollution	County	County	1-1	-
Air Pollution	City	City	1-1	-
Air Pollution	Date	Date	1-1	-
Air Pollution	O3 AQI	O3 AQI	1-1	-
Air Pollution	CO AQI	CO AQI	1-1	-
Air Pollution	SO2 AQI	SO2 AQI	1-1	-
Air Pollution	NO2 AQI	NO2 AQI	1-1	-
US Traffic	Severity	Severity	1-1	-
US Traffic	Start Time	Start Time	1-1	Conversion to the format YYYY-MM-DD
US Traffic	City	City	1-1	-
US Traffic	State	State	1-1	-
US Traffic	Amenity	Amenity	1-1	-
US Traffic	Bump	Bump	1-1	-
US Traffic	Crossing	Crossing	1-1	-
US Traffic	Give_Way	Give_Way	1-1	-
US Traffic	Junction	Junction	1-1	-
US Traffic	No Exit	No Exit	1-1	-
US Traffic	Railway	Railway	1-1	-
US Traffic	Roundabout	Roundabout	1-1	-
US Traffic	Station	Station	1-1	-
US Traffic	Stop	Stop	1-1	-
US Traffic	Traffic_Calming	Traffic_Calming	1-1	-
US Traffic	Traffic_Signal	Traffic_Signal	1-1	-
NBA Teams	abbreviation	Abbreviation	1-1	-
NBA Teams	city	city	1-1	-
NBA Teams	state	state	1-1	-

Figure 9: Correspondences between the datasets and the integrated model

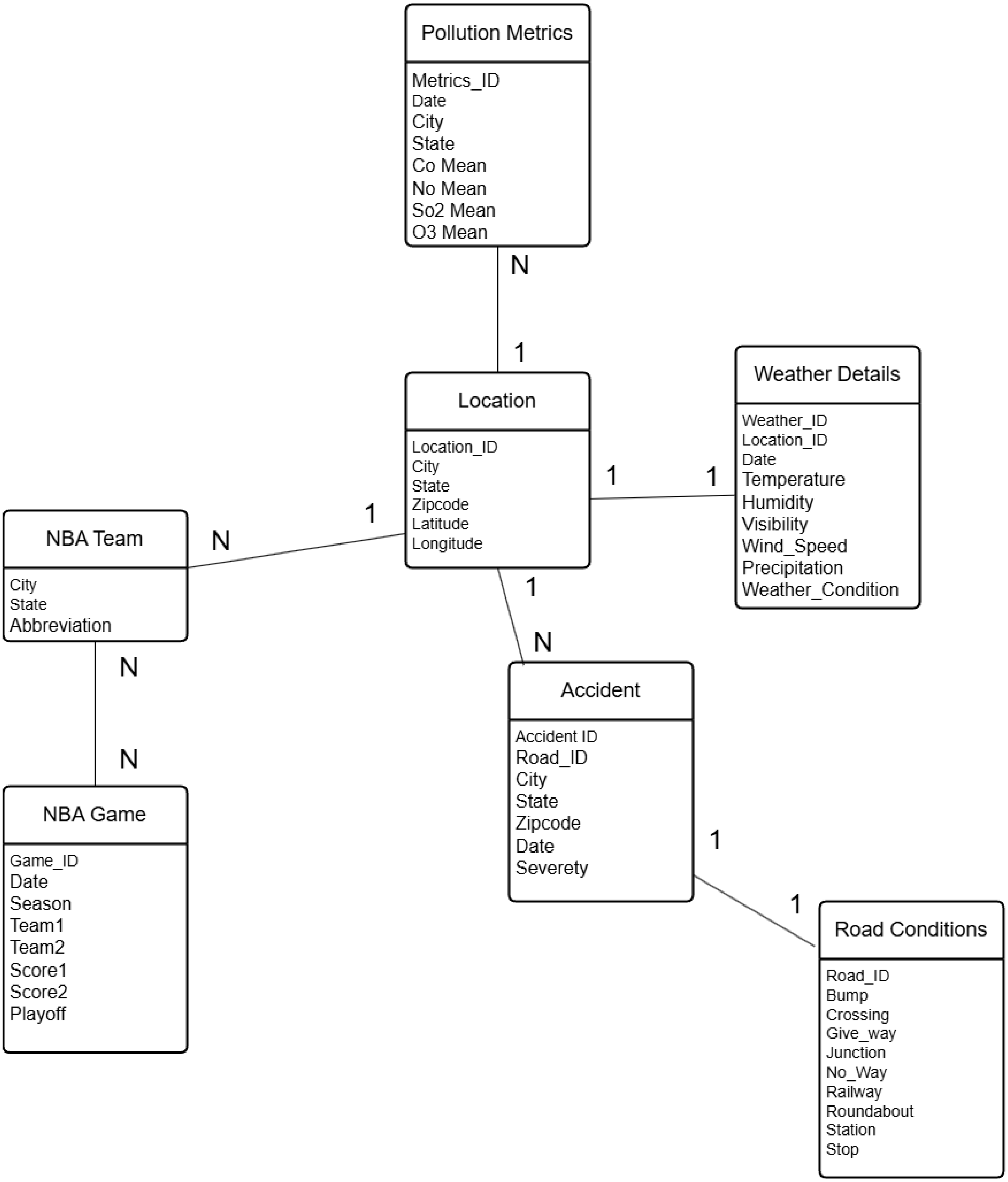


Figure 10: Entity-Relationship diagram of the datasets.

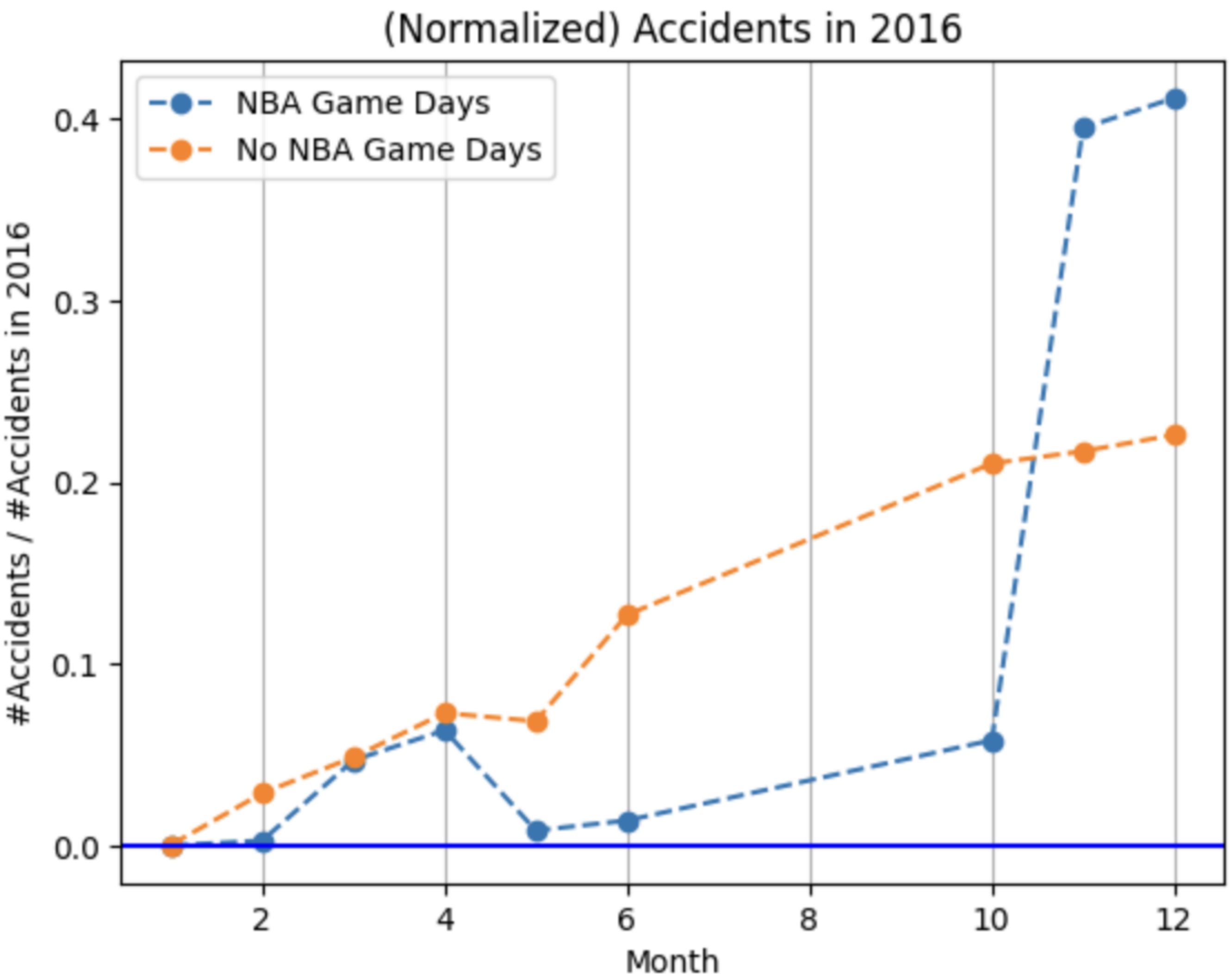


Figure 11: Monthly accident trends for 2016 by game loss and non-game days.

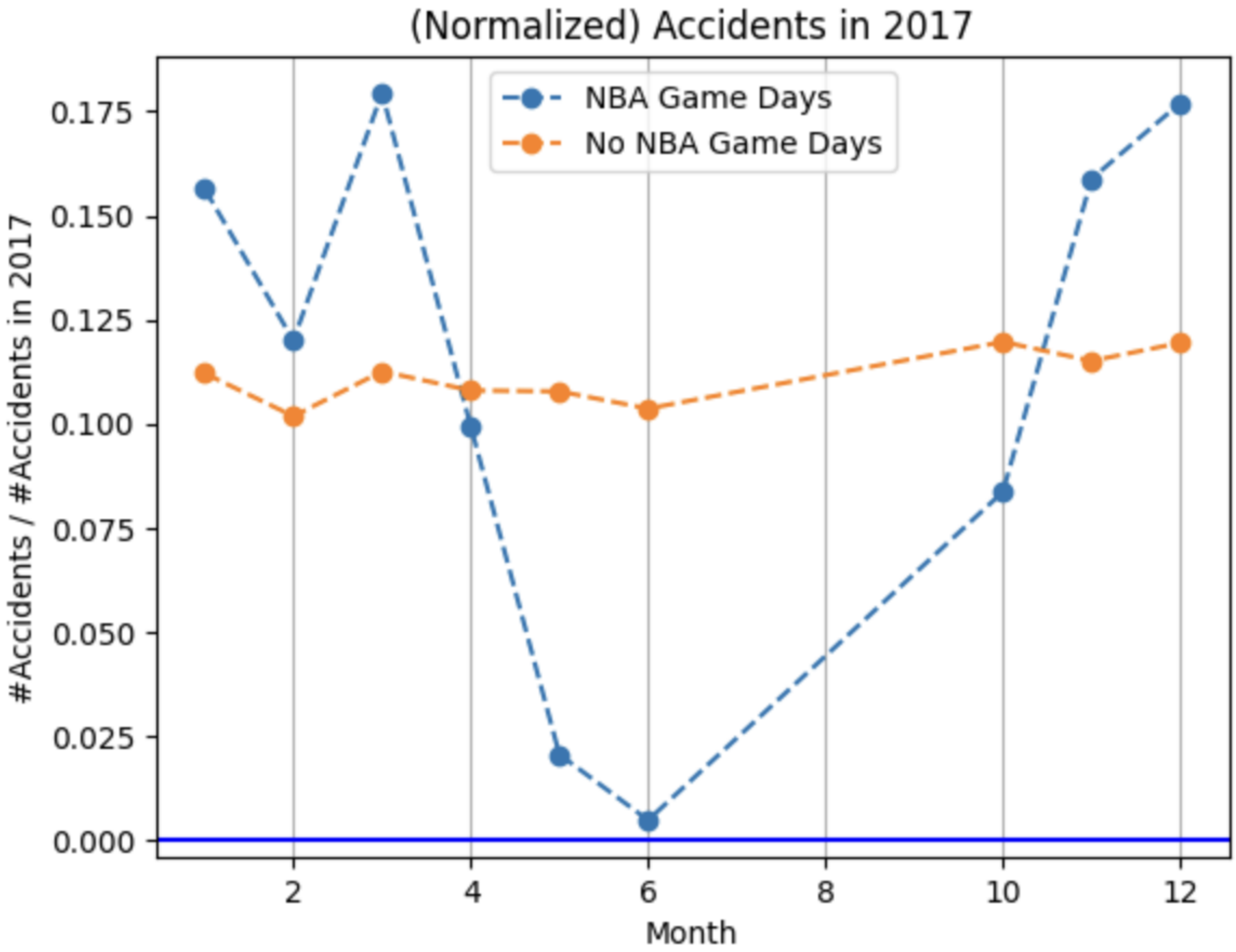


Figure 12: Monthly accident trends for 2017 by game loss and non-game days.

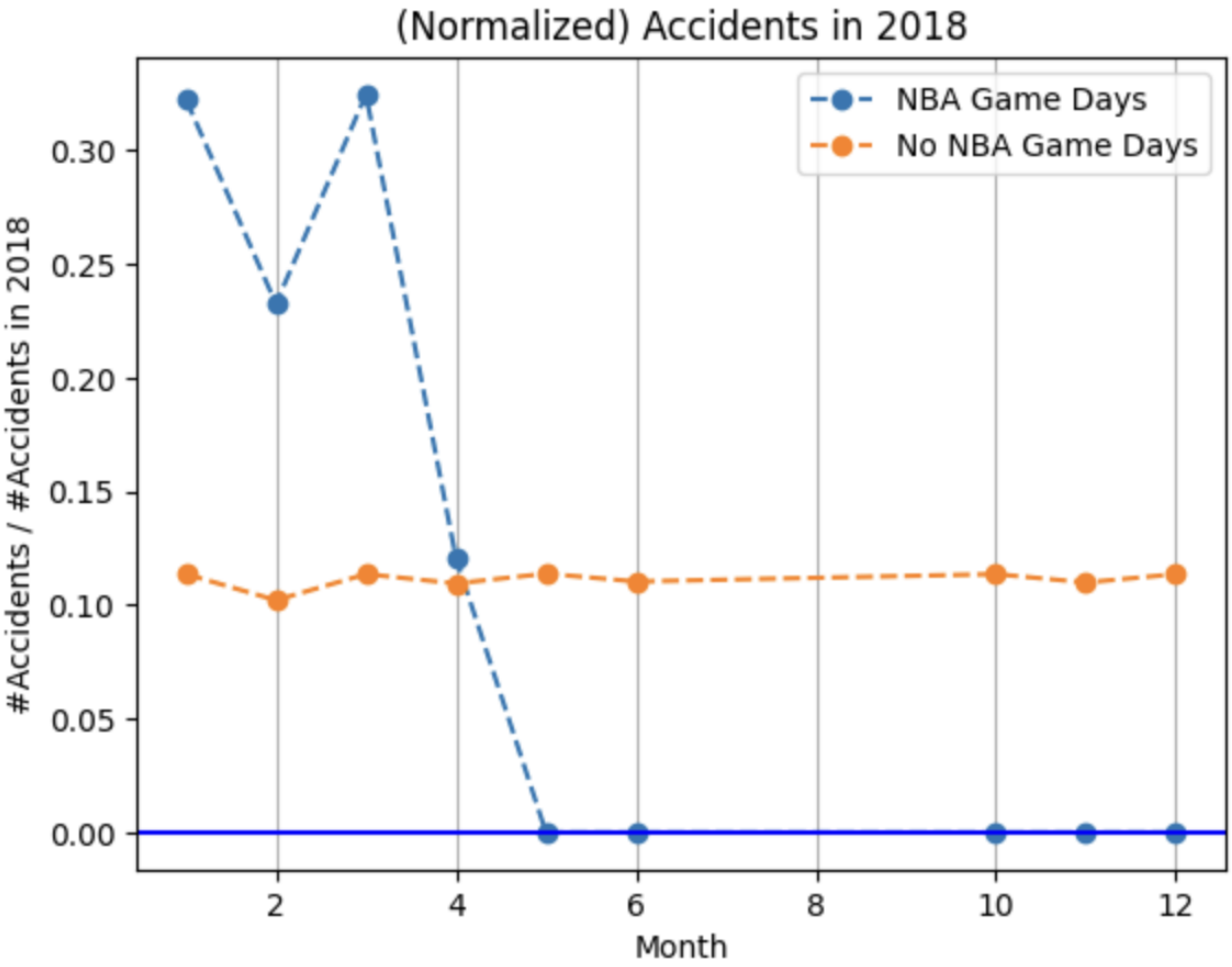


Figure 13: Monthly accident trends for 2018 by game loss and non-game days.

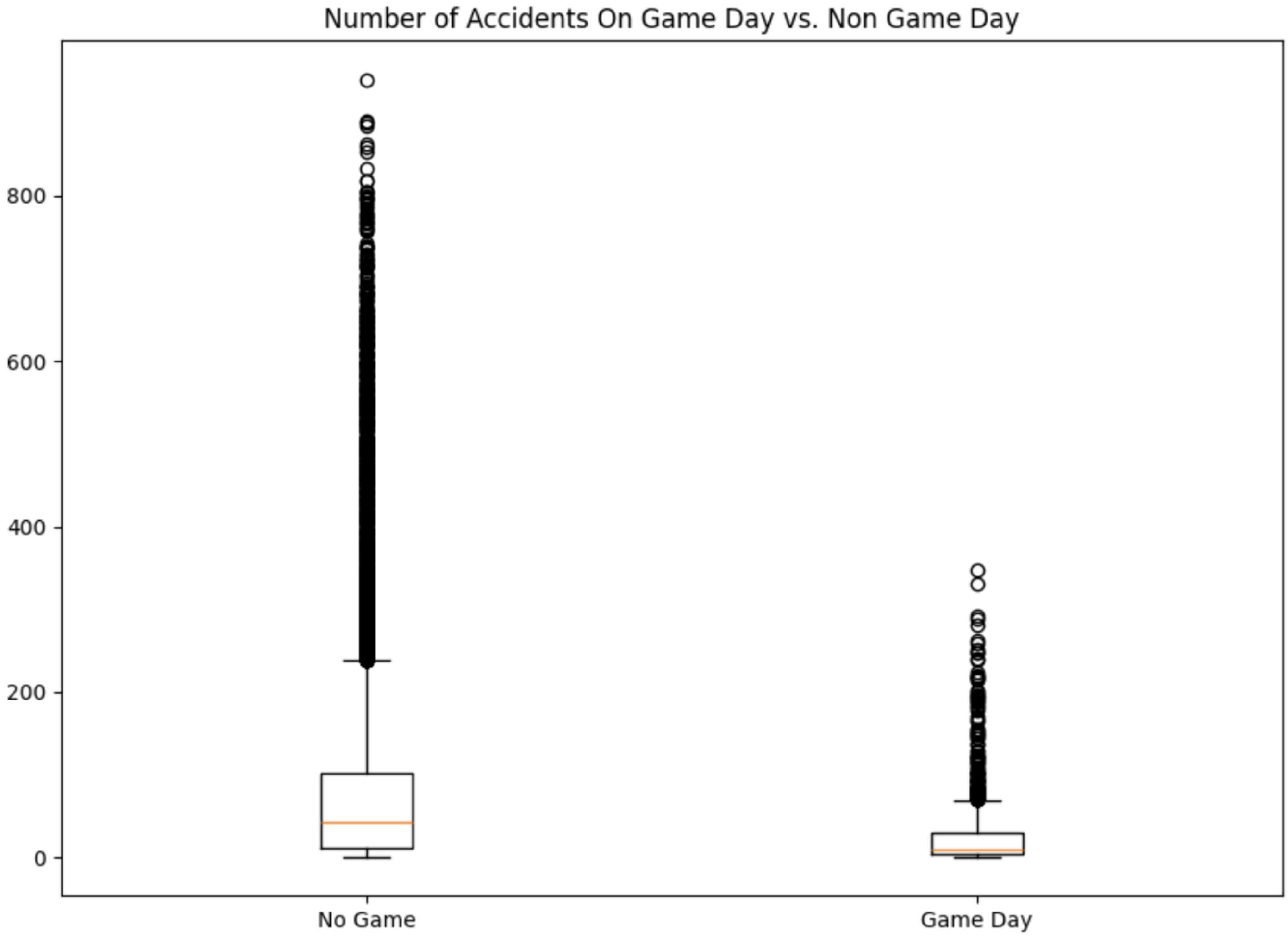


Figure 14: Accident count distribution per state on game loss and non-game days.

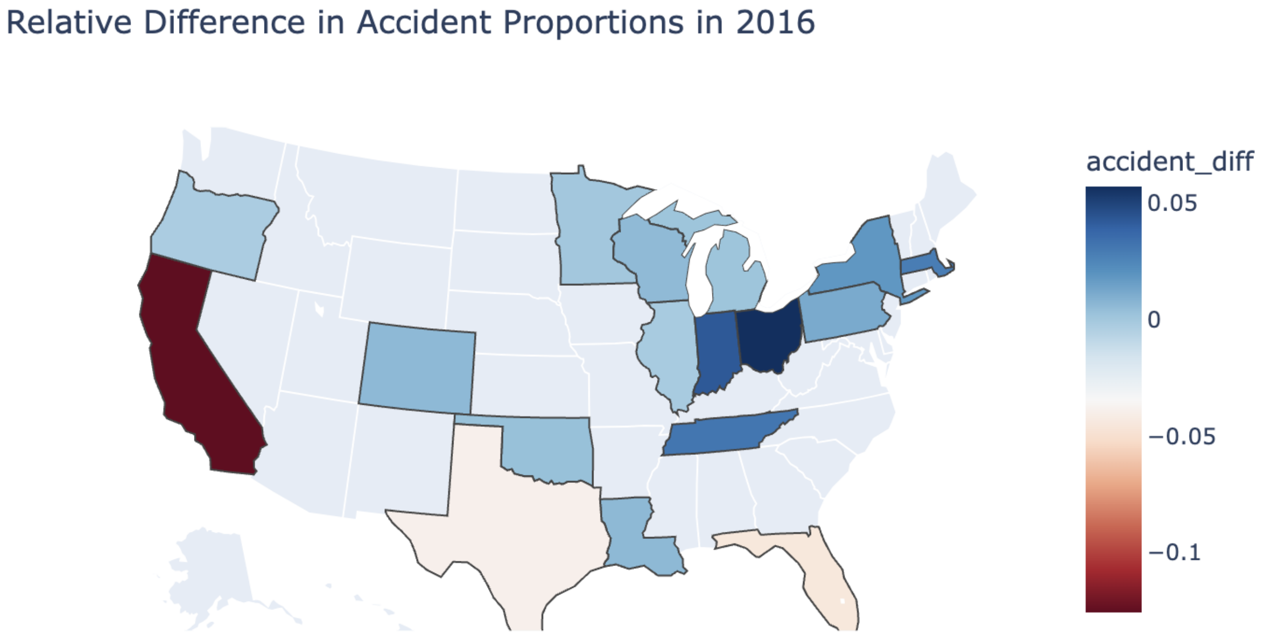


Figure 15: State-level proportion difference in accidents on game loss vs. non-game days in 2016.

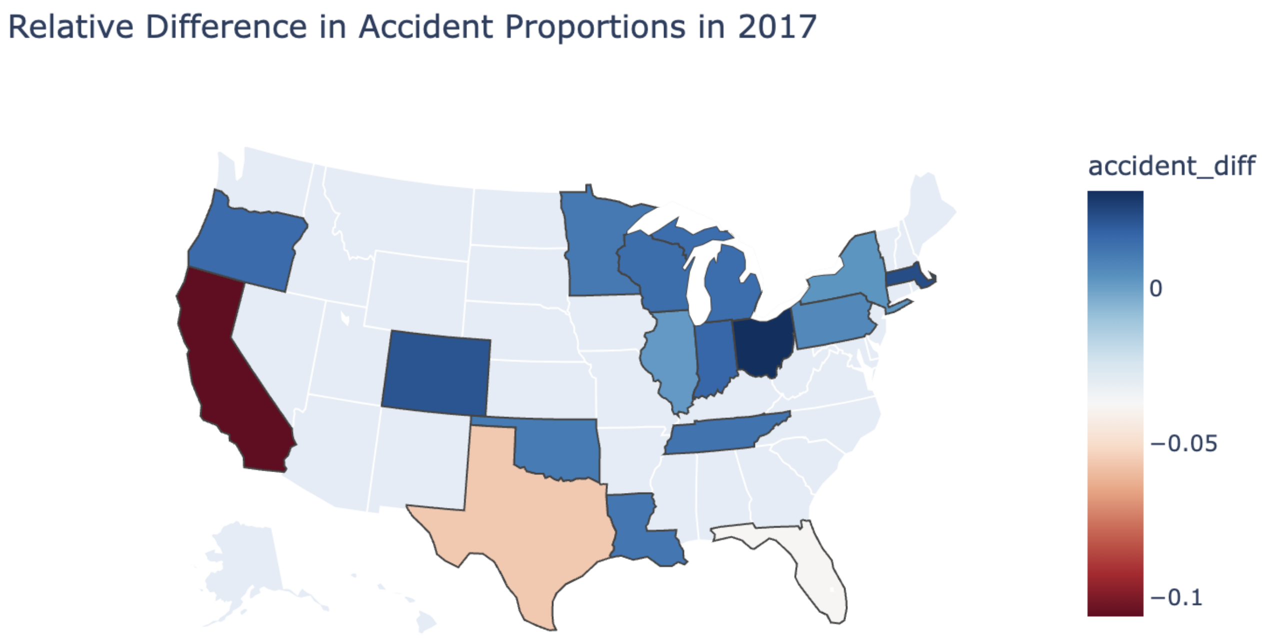


Figure 16: State-level proportion difference in accidents on game loss vs. non-game days in 2017.

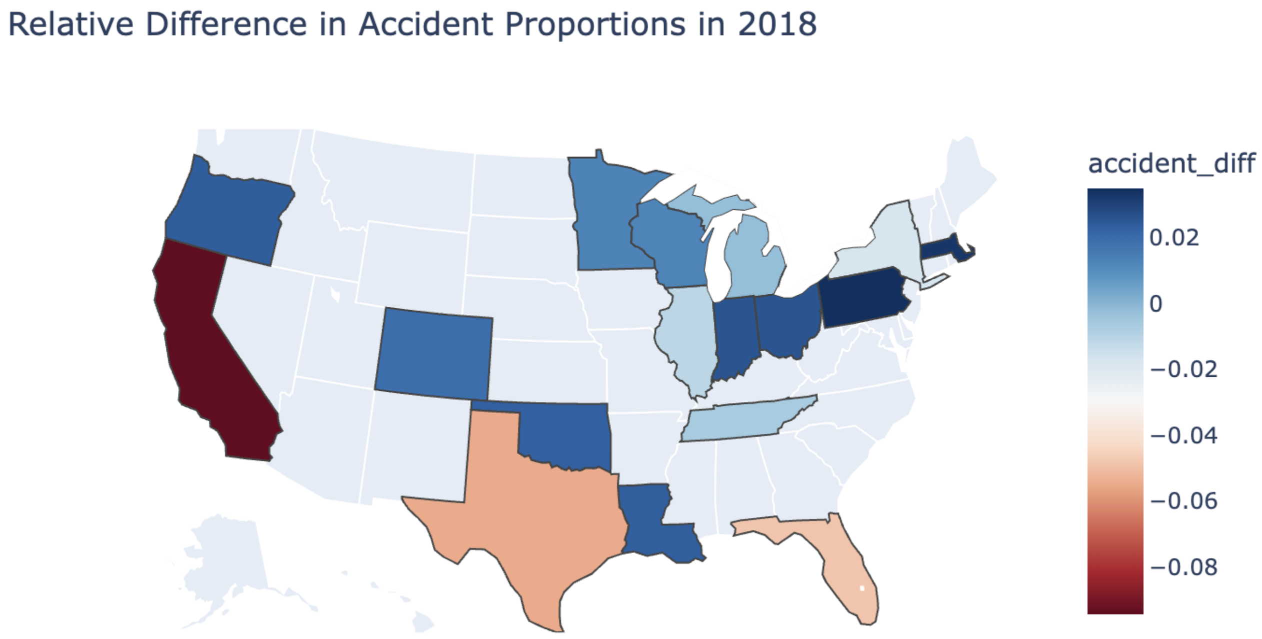


Figure 17: State-level proportion difference in accidents on game loss vs. non-game days in 2018.

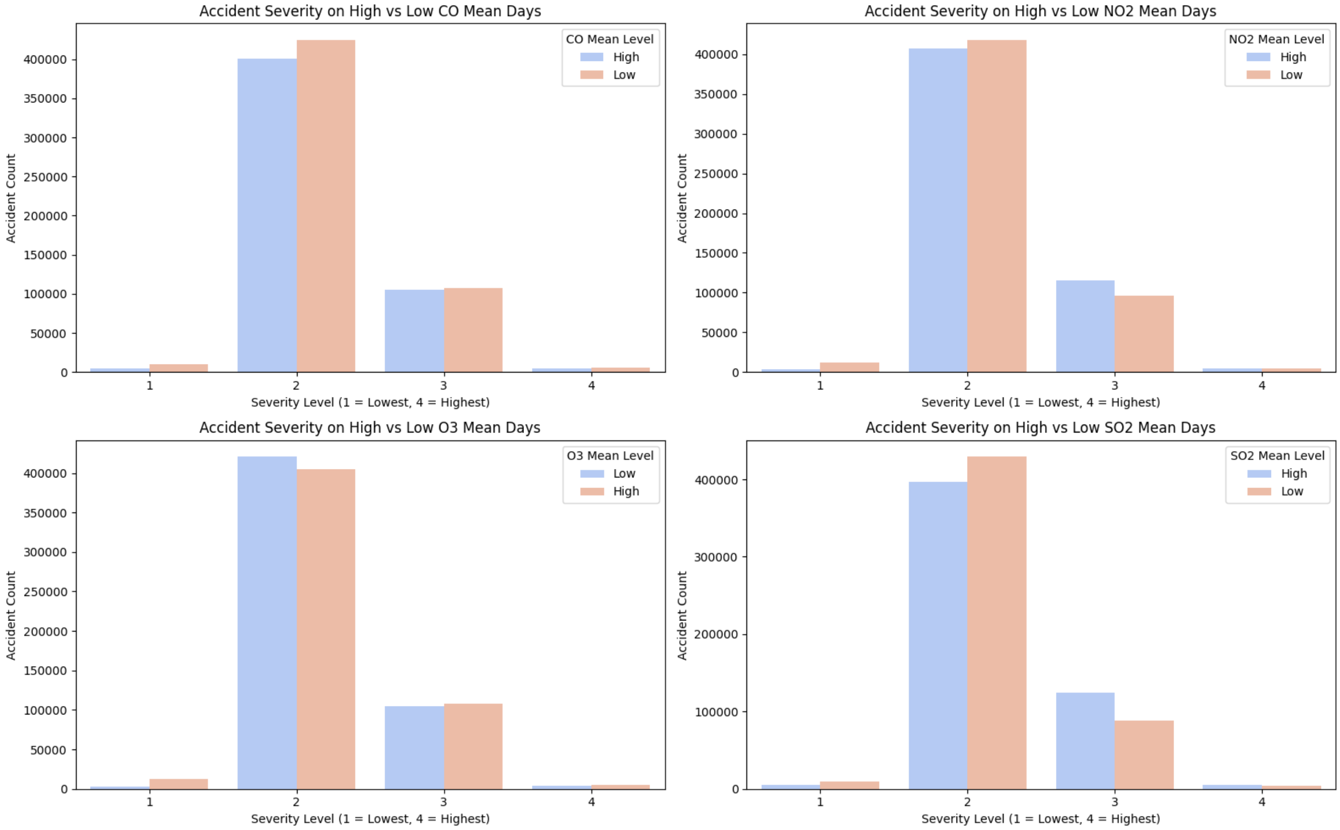


Figure 18: Severity distribution on high vs. low pollution days by pollutant.

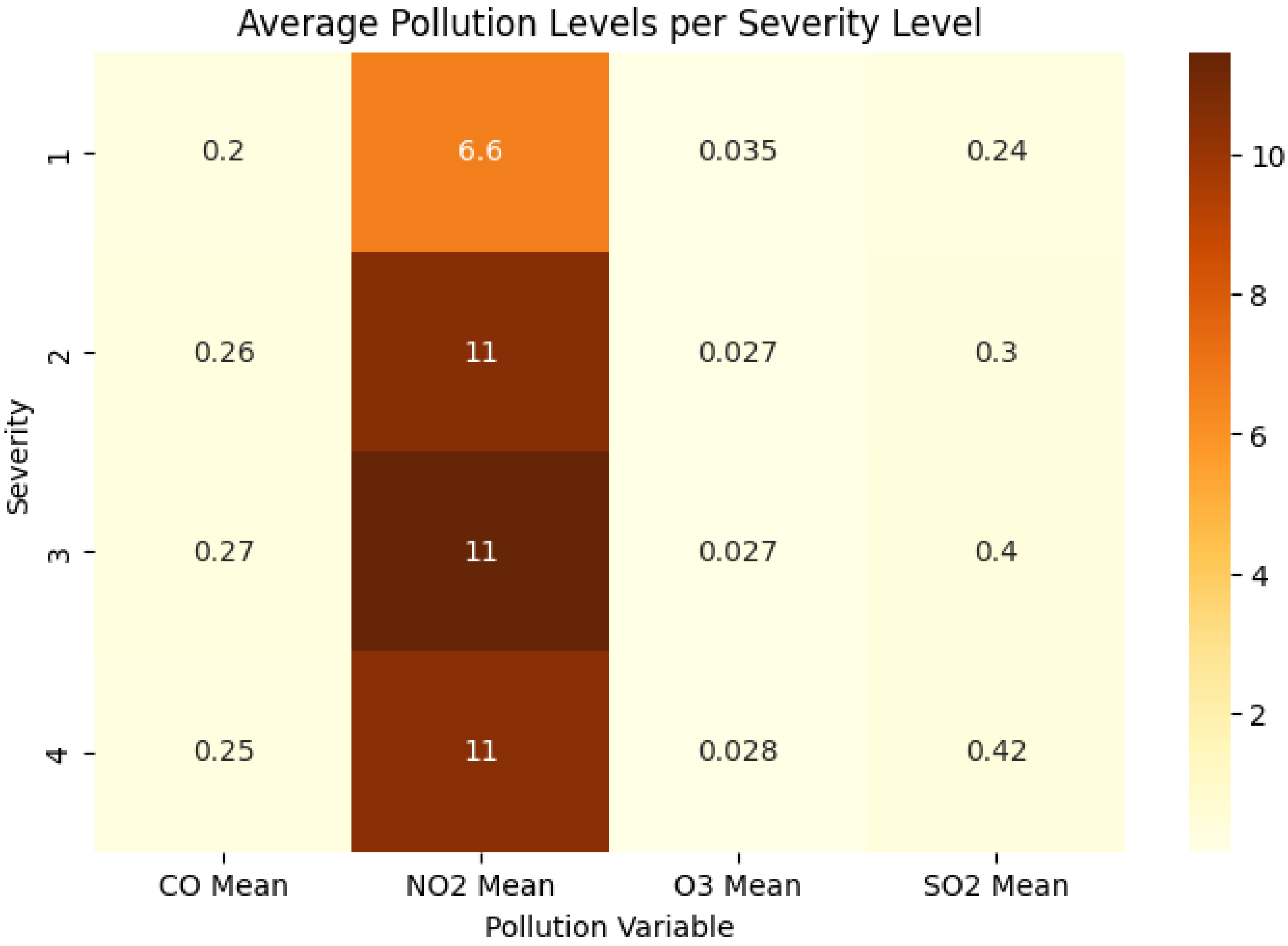
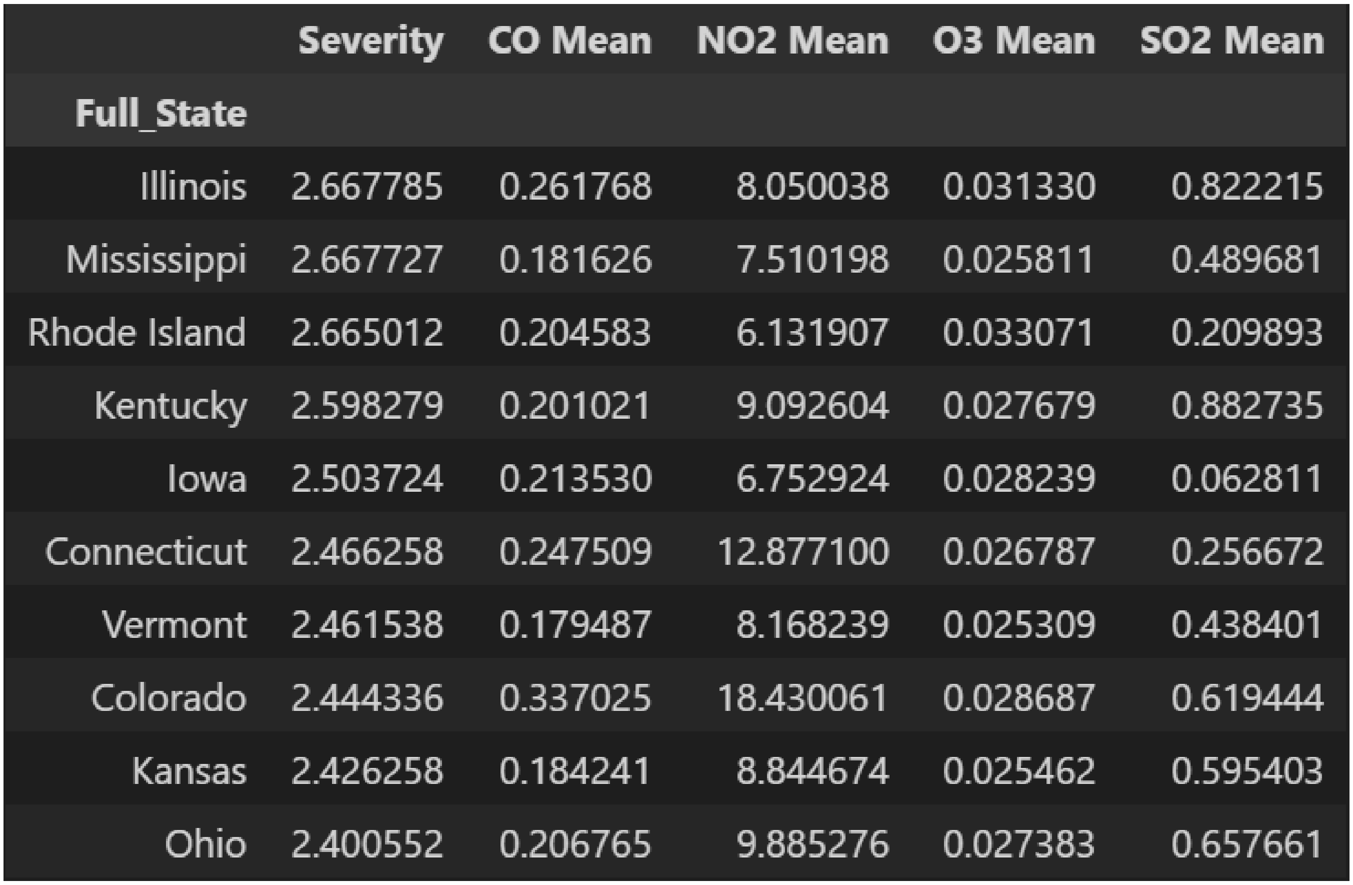
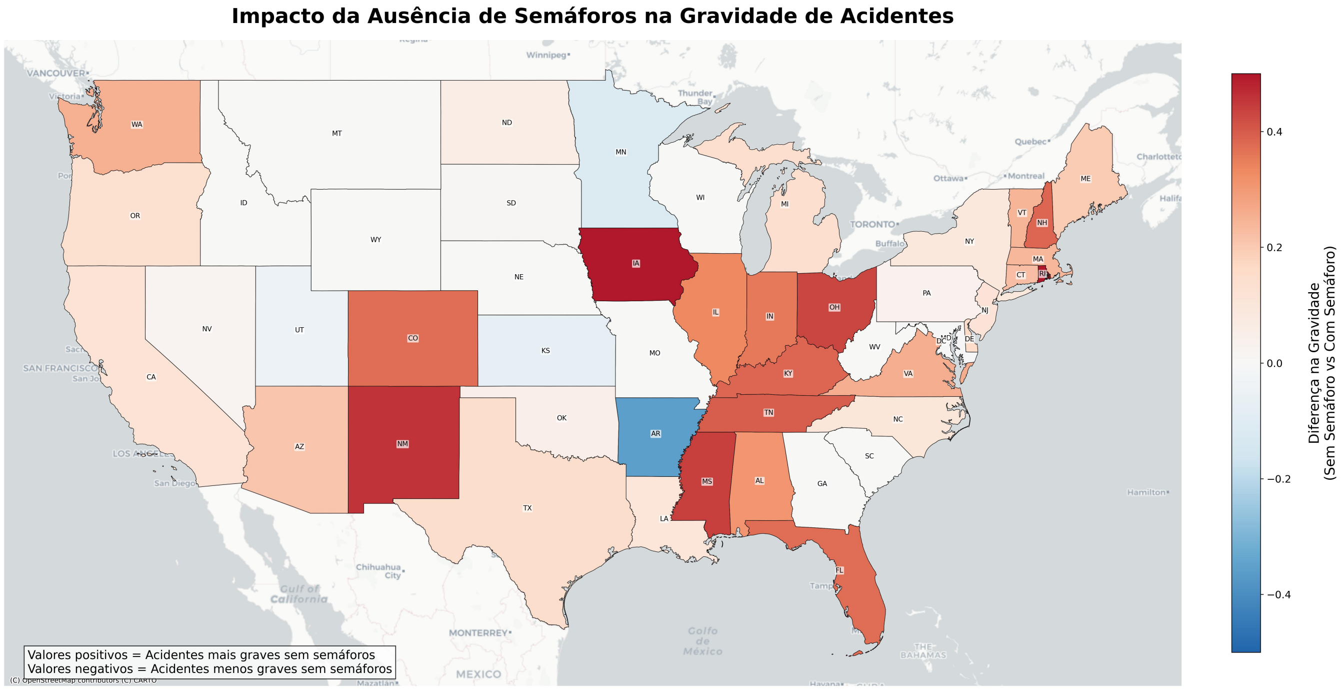


Figure 19: Average pollutant levels across severity levels.



	Severity	CO Mean	NO2 Mean	O3 Mean	SO2 Mean
Full_State					
Illinois	2.667785	0.261768	8.050038	0.031330	0.822215
Mississippi	2.667727	0.181626	7.510198	0.025811	0.489681
Rhode Island	2.665012	0.204583	6.131907	0.033071	0.209893
Kentucky	2.598279	0.201021	9.092604	0.027679	0.882735
Iowa	2.503724	0.213530	6.752924	0.028239	0.062811
Connecticut	2.466258	0.247509	12.877100	0.026787	0.256672
Vermont	2.461538	0.179487	8.168239	0.025309	0.438401
Colorado	2.444336	0.337025	18.430061	0.028687	0.619444
Kansas	2.426258	0.184241	8.844674	0.025462	0.595403
Ohio	2.400552	0.206765	9.885276	0.027383	0.657661

Figure 20: Top 10 states ranked by average severity and pollution levels.


Impacto da Ausência de Semáforos na Gravidade de Acidentes

Valores positivos = Acidentes mais graves sem semáforos
Valores negativos = Acidentes menos graves sem semáforos

(C) OpenStreetMap contributors (C) CARTO

Figure 21: Map of the absence of traffic lights in the severity of accidents.

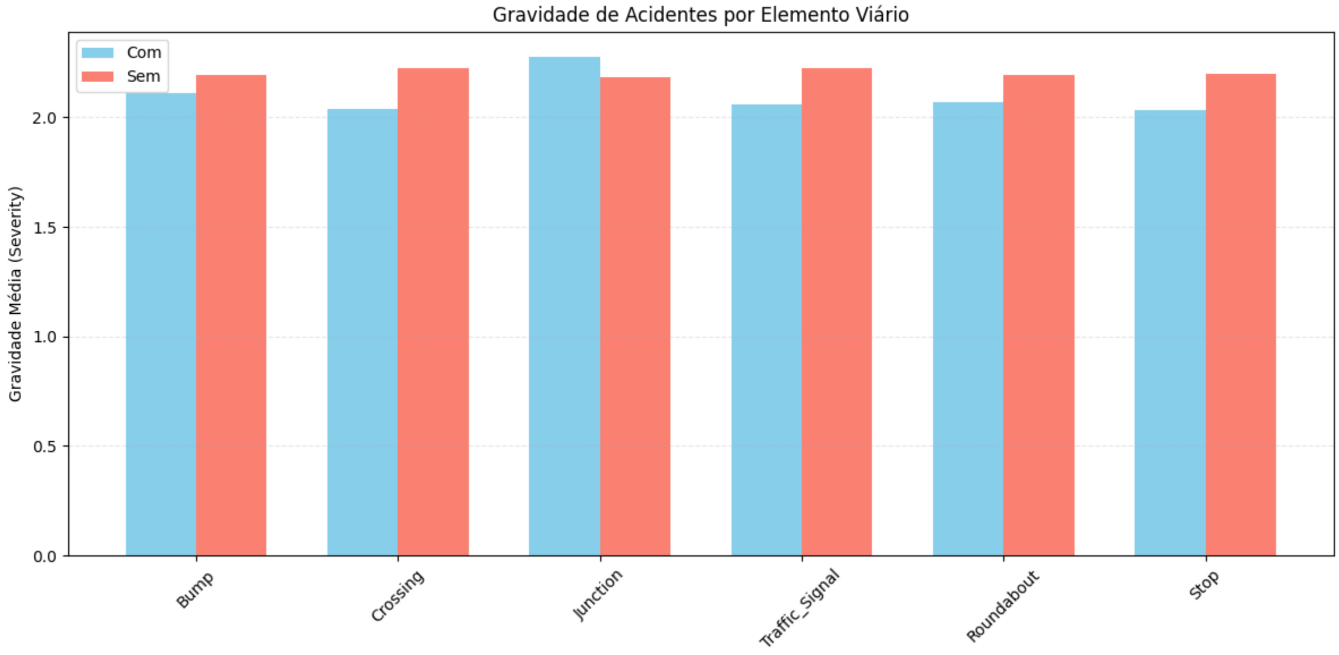

Gravidade de Acidentes por Elemento Viário

Figure 22: Histograms of accidents severity with vs. without road elements.