Project Phase 1 - 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 [6].

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 timeframes 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 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.

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

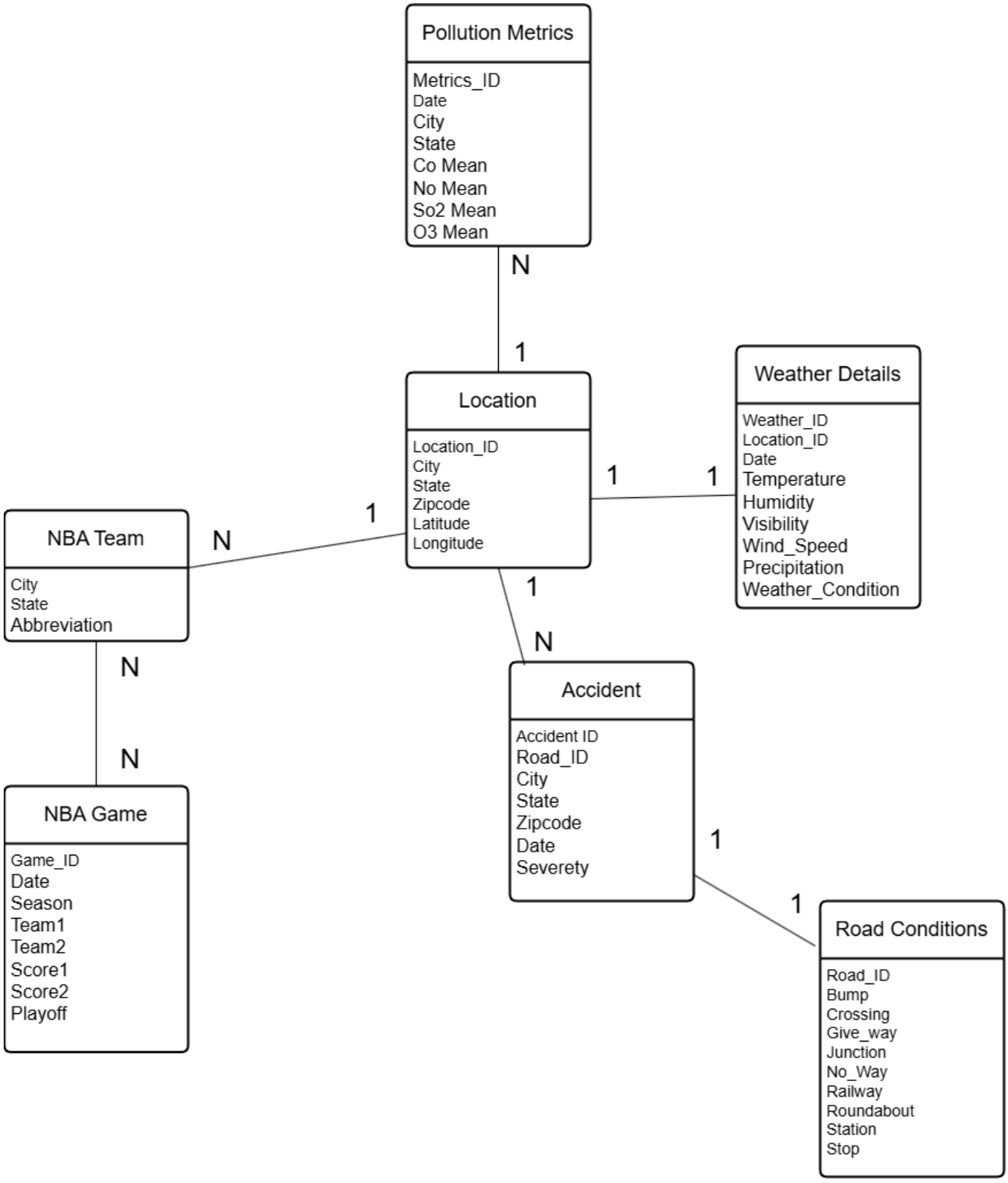

Figure 10: Entity-Relationship diagram of the datasets.