Methods of Advanced Data Engineering (MADE)

Data Report, Ismail Halil Kuzu

June 6, 2024

1 Question

How do different weather conditions effect the frequency and severity of traffic accidents in NYC?

2 Data Sources

I selected this datasets because it provides comprehensive coverage of traffic accidents and weather conditions in New York City for the year 2020. The detailed information on each accident is essential for my analysis of how weather conditions impact accident frequency and severity. The richness of this datas allows me to conduct a nuanced analysis of contributing factors and trends over time.

Datasource1: NYC Traffic Accidents

- **Metadata URL:** NYC Traffic Accidents - Metadata
- **Data URL:** NYC Traffic Accidents - Data
- **Data Type:** CSV

This data source contains NYC Traffic Accident details for the year 2020. The dataset includes several attributes such as CRASH DATE and CRASH TIME, BOROUGH, ON STREET NAME, LOCATION and other relevant attributes, which will be considered for analytics.

Datasource2: Weather Data for NYC

- **Metadata URL:** NYC Weather - Metadata
- **Data URL:** NYC Weather - Data
- **Data Type:** CSV

This dataset provides a comprehensive overview of weather conditions in New York City spanning from 2016 to 2022. It includes attributes such as TIME, temperature at 2 meters above ground level (temperature 2m in °C), precipitation (mm), rain (mm), cloud cover percentage , low-level cloud cover percentage, mid-level cloud cover percentage, high-level cloud cover percentage, wind speed at 10 meters above ground level (windspeed 10m in km/h), and wind direction at 10 meters above ground level (winddirection 10m in °). For our analysis, we will focus specifically on the data from the year 2020 to extract insights and patterns relevant to that year.

CRASH DATE	CRASH TIME	BOROUGH	ZIP CODE	LATITUDE	LONGITUDE	LOCATION	ON STREET NAME	CROSS STREET NAME	OFF STREET NAME	NUMBER OF PERSONS INJURED
2020-08-29	15:40:00	BRONX	10466	40.8921	-73.83376	POINT (-73.83376 40.8921)	PRATT AVENUE	STRANG AVENUE		0
2020-08-29	21:00:00	BROOKLYN	11221	40.6905	-73.919914	POINT (-73.919914 40.6905)	BUSHWICK AVENUE	PALMETTO STREET		2
2020-08-29	18:20:00			40.8165	-73.946556	POINT (-73.946556 40.8165)	8 AVENUE			1
2020-08-29	00:00:00	BRONX	10459	40.82472	-73.89296	POINT (-73.89296 40.82472)			1047 SIMPSON STREET	0
2020-08-29	17:10:00	BROOKLYN	11203	40.64989	-73.93389	POINT (-73.93389 40.64989)			4609 SNYDER AVENUE	0
2020-08-29	03:29:00			40.68231	-73.84495	POINT (-73.84495 40.68231)	WOODHAVEN BOULEVARD			1
2020-08-29	19:30:00	BRONX	10459	40.825226	-73.88778	POINT (-73.88778 40.825226)	LONGFELLOW AVENUE	EAST 165 STREET		0
2020-08-29	00:00:00			40.80016	-73.93538	POINT (-73.93538 40.80016)	2 AVENUE			0
2020-08-29	19:50:00	BRONX	10466	40.894314	-73.86027	POINT (-73.86027 40.894314)	EAST 233 STREET	CARPENTER AVENUE		0
2020-08-29	09:20:00	QUEENS	11385	40.70678	-73.90888	POINT (-73.90888 40.70678)			565 WOODWARD AVENUE	0
2020-08-29	00:07:00	QUEENS	11436	40.680237	-73.79774	POINT (-73.79774 40.680237)			116-52 144 STREET	0
2020-08-29	14:00:00	QUEENS	11433	40.704422	-73.792854	POINT (-73.792854 40.704422)	ARCHER AVENUE	MERRICK BOULEVARD		0
2020-08-29	21:33:00	BRONX	10455	40.812965	-73.9161	POINT (-73.9161 40.812965)	EAST 146 STREET	BROOK AVENUE		1
2020-08-29	22:53:00	BROOKLYN	11249	40.70166	-73.961464	POINT (-73.961464 40.70166)	WILLIAMSBURG STREET WEST	WYTHE AVENUE		0
2020-08-29	04:14:00			40.835373	-73.842186	POINT (-73.842186 40.835373)	WATERBURY AVENUE			1
2020-08-29	06:35:00			40.65965	-73.773834	POINT (-73.773834 40.65965)	ROCKAWAY BOULEVARD	NASSAU EXPRESSWAY		0
2020-08-29	13:00:00	BROOKLYN	11206	40.699707	-73.95718	POINT (-73.95718 40.699707)	BEDFORD AVENUE	WALLABOUT STREET		0
2020-08-29	10:30:00	QUEENS	11385	40.7122	-73.86208	POINT (-73.86208 40.7122)	METROPOLITAN AVENUE	COOPER AVENUE		2
2020-08-29	12:29:00	BRONX	10453	40.861862	-73.91282	POINT (-73.91282 40.861862)	WEST FORDHAM ROAD	MAJOR DEEGAN EXPRESSWAY		2
2020-08-29	10:35:00	BROOKLYN	11211	40.710957	-73.951126	POINT (-73.951126 40.710957)	UNION AVENUE	GRAND STREET		1
2020-08-29	13:55:00	BROOKLYN	11231	40.67473	-74.00029	POINT (-74.00029 40.67473)	HAMILTON AVENUE	GARNET STREET		1
2020-08-29	00:30:00			40.66584	-73.75551	POINT (-73.75551 40.66584)	BELT PARKWAY			0
2020-08-29	06:30:00			40.65052	-73.73309	POINT (-73.73309 40.65052)	CRAFT AVENUE			0
2020-08-29	19:00:00			40.83968	-73.929276	POINT (-73.929276 40.83968)	MAJOR DEEGAN EXPRESSWAY			1
2020-08-29	01:45:00	MANHATTAN	10029	40.79477	-73.93247	POINT (-73.93247 40.79477)			545 EAST 116 STREET	0
2020-08-29	08:45:00	QUEENS	11411	40.701042	-73.74636	POINT (-73.74636 40.701042)			114-52 208 STREET	0
2020-08-29	23:19:00	BROOKLYN	11226	40.63962	-73.95477	POINT (-73.95477 40.63962)	NEWKIRK AVENUE	FLATBUSH AVENUE		1
2020-08-29	07:10:00			40.674347	-73.82071	POINT (-73.82071 40.674347)	118 STREET			0
2020-08-29	00:56:00	BRONX	10461	40.84307	-73.848076	POINT (-73.848076 40.84307)	EAST TREMONT AVENUE	SILVER STREET		1

Figure 1: Raw accident data from the data source

Data Structure and Quality: Both datasets are set up like tables and come in CSV format, making them easy to read and work with. In the accident dataset, the first row tells you what each column is about. In the weather datasets, the information is organized by dates, so you can easily see what the weather was like on any given day.

Accuracy: The data sets I have chosen are collected from authentic sources. So, these data reflect the real world data and 100 percent correctness is ensured.

Consistency: All the data sets are in tabular format. So the data format is relevant for all kinds of places.

Relevancy: Both data sets are considered for year 2020.

Validity checks: The number of missing, invalid, and duplicate values in the columns has been cleaned from both data sets.

Data sources licenses obligations: Accident and weather datasets have been collected from public sources and are available for free use.

For NYC Traffic Accidents dataset there is no copyright issues. NYC Weather - 2016 to 2022 dataset provides weather data licensed under non-commercial terms (CC BY-NC 4.0), allowing for sharing and non-commercial use.

Weather Datasets License Terms: https://creativecommons.org/licenses/by-nc/4.0/
Accident Datasets License Terms: https://creativecommons.org/publicdomain/zero/1.0/

So, to follow the data obligations, I will refrain from using the data for any commercial purposes, ensuring that all usage remains within the scope of non-commercial activities.

Data Pipeline

Technology Used: I used Python, along with libraries like `requests`, `zipfile`, `pandas`, and `sqlite3`, to develop the data pipeline.

Data Transformation Steps:

- First, I fetched the datasets from the provided URLs using the `requests` library.
- Then, I extracted the downloaded zip files to the designated directory to access the CSV files.
- After that, I read the CSV files into pandas dataframes for further processing.
- To handle missing values, I filled them with zeros to maintain consistency across the datasets.
- Finally, I stored the processed dataframes as tables in a SQLite database named `nyc_climate_traffic.db`.

Problems Encountered and Error Handling: I faced challenges during the decompression of the zip files, especially with the weather data. I resolved this issue by utilizing the `zipfile` library for extraction. Additionally, I implemented robust error-handling mechanisms to manage exceptions, such as corrupted zip files. If any errors occur, the system notifies me with an error message.

Result and Limitations

Data Output: The output of my data pipeline is a SQLite database (`nyc_climate_traffic.db`) containing two tables: `traffic_accidents` and `weather_data`. These tables hold the cleaned and processed data from the respective datasets.

Data Structure and Quality: The output maintains the structure and quality of the input data, with missing values handled appropriately. I chose SQLite as the output format due to its seamless integration with various analysis tools and efficient data retrieval capabilities.

Limitations and Potential Issues: Despite successful processing, outliers or anomalies may exist in the data. Therefore, rigorous data profiling and validation techniques are essential to identify and address any issues before analysis. Additionally, changes in input data structures or formats may require adjustments to the pipeline to ensure continued functionality.