

Data Report

processed by Prosenjit Chowdhury (23361276)

1.1 Question

What is the impact of the weather and climate conditions in New York with its road traffic volume?

1.2 Data Sources

Data source-01: Weather Data of the City of New York

Metadata URL: <https://meteostat.net/en/place/us/new-york-city?s=72502&t=2012-01-01/2012-12-31>

Sample Data: <https://bulk.meteostat.net/v2/hourly/72502.csv.gz>

These data sets were selected to determine how a city's weather affects its traffic volume. New York is one of the largest and busiest cities in the world that I consider. So, I collected data on the weather and traffic volume of that city. The year 2012 will be the subject of analysis. Weather for that year will be examined, and its effect on traffic volume based on that year will also be examined.

Data Type: CSV

Data source-02: Traffic Volume Data of the City of New York

Metadata URL: <https://catalog.data.gov/dataset/traffic-volume-counts>

Sample Data: <https://data.cityofnewyork.us/api/views/btm5-ppia/rows.csv>

This data source features traffic accidents in New York City from 1974 to 2024 Only the year 2012 will be considered for the analysis of all years There are several features like date, time, street name, directions, etc.

Data Type: CSV

ID	Segment	Roadway	From	To	Direction	Date	12:00-1:00	1:00-2:00	2:00-3:00	3:00-4:00	4:00-5:00	5:00-6:00	6:00-7:00	7:00-8:00	8:00-9:00
1	15540	BEACH ST	UNION PL	VAN DUZENB		1/9/2012	20	10	11	14	13	20	34	66	100
2	15540	BEACH ST	UNION PL	VAN DUZENB		1/10/2012	21	16	8	6	13	13	31	70	67
3	15540	BEACH ST	UNION PL	VAN DUZENB		1/11/2012	27	14	6	5	12	16	34	75	69
4	15540	BEACH ST	UNION PL	VAN DUZENB		1/12/2012	22	7	7	8	11	12	33	75	89
5	15540	BEACH ST	UNION PL	VAN DUZENB		1/13/2012	31	17	7	5	13	28	29	68	84
6	15540	BEACH ST	UNION PL	VAN DUZENB		1/14/2012	42	27	21	18	21	13	17	18	46
7	15540	BEACH ST	UNION PL	VAN DUZENB	SB	1/9/2012	27	12	12	4	22	27	66	154	155
8	15540	BEACH ST	UNION PL	VAN DUZENB	SB	1/10/2012	26	16	11	13	16	27	59	156	177
9	15540	BEACH ST	UNION PL	VAN DUZENB	SB	1/11/2012	32	16	8	9	15	26	63	169	178
10	15540	BEACH ST	UNION PL	VAN DUZENB	SB	1/12/2012	24	12	7	18	11	23	61	146	177
11	15540	BEACH ST	UNION PL	VAN DUZENB	SB	1/13/2012	39	22	8	6	16	30	77	147	187
12	15540	BEACH ST	UNION PL	VAN DUZENB	SB	1/14/2012	45	45	28	23	21	19	32	56	67
13	12809	LITTLE CLC	NORTHER	BRISTOL A	EB	1/9/2012	21	11	2	3	11	29	117	511	328
14	12809	LITTLE CLC	NORTHER	BRISTOL A	EB	1/10/2012	20	14	5	2	13	24	104	477	322
15	12809	LITTLE CLC	NORTHER	BRISTOL A	EB	1/11/2012	16	11	3	6	12	24	118	519	348
16	12809	LITTLE CLC	NORTHER	BRISTOL A	EB	1/12/2012	26	14	7	1	6	27	128	483	390

Figure 01: Raw traffic volume data from the data source

Data Structure & Quality: The traffic volume dataset is organized tabularly, with columns indicating road locations within New York City and total volumes during specific months. However, the hourly weather reports in the Weather Dataset are similarly organized in a tabular format. The format of both datasets is CSV.

Accuracy: The above data set is collected from a trusted and widely used website thereby indicating that the respective data set is completely accurate and reliable.

Consistency: The same data format is replicated at each location to maintain data consistency and quality.

Relevancy: Both data sets consider data from 2012 so that only relevant data events and complete weather data from that year are appropriate.

Validity checks: After both data sets were thoroughly analyzed, the columns cleared for final use. The missing, invalid, and duplicate member value data from the columns were removed.

Data sources licenses & obligations: Weather and traffic volume number datasets are freely accessible and collected from open sources.

Terms and conditions: Users may use the City of New York dataset non-commercially under the terms of an Open Data License. **Meteostat** makes weather data available under the CC BY-NC 4.0 Non-Commercial License, which allows sharing and non-commercial use.

Terms of Service for Meteo Dataset License: <https://dev.meteostat.net/terms.html#use-of-services>

License terms for traffic volume number dataset: <https://resources.data.gov/open-licenses/>

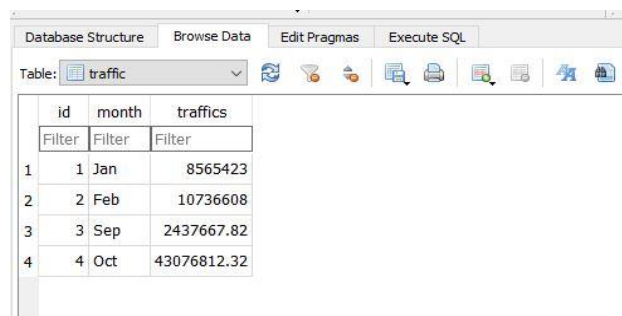
I will therefore refrain from using the data for any commercial purposes in order to comply with data obligations, ensuring that all uses remain within the scope of non-commercial activities.

1.3 Data Pipeline

Technology Details: Python packages such as Pandas for data manipulation and SQLite for data storage are used to implement the data pipeline.

Data transformation and cleaning tread:

- Using the requests library, both data sets are retrieved from their respective URLs.
- To provide a monthly total of traffic volume, traffic datasets are modified.
- Weather data for 2012 is sorted and grouped by month so that the monthly averages can be examined.
- Ultimately, a SQLite database is used to store both data tables for future study.



	id	month	traffics
1	1	Jan	8565423
2	2	Feb	10736608
3	3	Sep	2437667.82
4	4	Oct	43076812.32

Figure 02: Modified traffic volume data from the data table.

Problems Encountered and Solutions

When retrieving weather data, the primary problem was handling compressed data. Using the gzip package to first decompress the data before processing helped remedy this.

It was quite labor-intensive to ensure data consistency and integrity, especially when dealing with missing or incorrect entries. Rigorous data cleaning and validation processes were used to correct these.

Error Handling and Input Data Changes

Error-handling procedures are established to detect and monitor exceptions that may occur during data processing.

The pipeline is designed to gracefully accommodate changing input data by dynamically adapting to new data structures or formats.

1.4 Result and Limitations:**Output data**

The output of the data pipeline is represented by two tables in the SQLite database: Weather, which contains monthly weather averages, and TrafficVolume, which contains monthly traffic volume number data for 2012. Both tables are organized with appropriate fields and data types to facilitate efficient interpretation and search.

Data Structure and Quality

By maintaining the accuracy and consistency of the stored data, the output maintains the quality and structure of the input data. Due to the wide array of output formats one can easily integrate SQLite with numerous frameworks and analysis tools.

Reflections and potential issues

At the analysis stage, even though the data pipeline processes and stores the data correctly, potential problems such as outliers or inconsistencies in the data may appear. Before drawing conclusions from the analysis, it is essential to carry out a comprehensive data search and validation to identify and resolve any problems of this nature.