Methods of Advanced Data Engineering (MADE)

Data Report

Report By: MD TANVER SADIK RAJ (Matriculation Number: 23359384)

1.1 Main Question

What is the major Impact of Weather & Climate condition on Average Daily Traffic Counts?

1.2 Data Sources

In my analysis and data sets my purpose was to find the impact of weather and climate condition on average daily traffic counts. In my data sets I have chosen data from the city Chicago for the year 2006. Also, for that same year I have also collected weather data, in order to identify the conditions. Based on the weather report of that year my goal is to identify the traffic counts of that city month wise so that we can observe a clear view how the weather conditions affect our day-to-day movement.

Data source for Weather data: Weather & Climate condition Data of the City Chicago for the year- 2006.

Metadata: https://meteostat.net/en/station/72534?t=2006-01-01/2006-12-31

Sample Data: https://bulk.meteostat.net/v2/hourly/72534.csv.gz

Data Type: CSV

This data source will provide weather and climate data in Chicago for the year 2006. This data set is generated from [Meteostat](https://meteostat.net/en/). This data set includes datetime, temp-max, temp-min, humidity, snow, windspeed, visibility, cloud cover etc.

Data source for Traffic Condition: Average Daily Traffic Counts Data in Chicago.

Metadata: https://catalog.data.gov/dataset/average-daily-traffic-counts/resource/fa81c305-4308-40ab-8f95-f7063fdcf769

Sample Data: https://data.cityofchicago.org/api/views/pfsx-4n4m/rows.csv

Data Type: CSV

This data set is for the city of Chicago and it's for the year 2006. In this specific year the average traffic count of the city is logged and there are various attributes present there to get a clear view of the data set. Some of the attributes are Traffic Volume Count Location Address, Street, Date, Total vehicle passing etc.

ID	Traffic Volume Count Location Address	Street	Date of Count	Total Passing Vehicle Volume	Longitude	Location	Vehicle Volume By Each Direction of Traffic
414	5838 West	Lake St	11/14/2006	7100	-87.771064	(41.887904, -87.771064)	East Bound: 3600 / West Bound: 3500
176	120 East	76th St	3/28/2006	8600	-87.617335	(41.756542, -87.617335)	East Bound: 3800 / West Bound: 4800
1367	1730 East	57th Dr	8/24/2006	53500	-87.582231	(41.792863, -87.582231)	East Bound: 27800 / West Bound: 25700
316	125 East	24th St	3/30/2006	700	-87.622658	(41.849302, -87.622658)	East Bound: 400 / West Bound: 300
1294	2924 East	130th St	8/29/2006	4200	-87.552112	(41.659177, -87.552112)	East Bound: 2300 / West Bound: 1900
507	1931 West	Lake St	10/19/2006	6900	-87.675556	(41.885023, -87.675556)	East Bound: 3400 / West Bound: 3500
691	6067 North	Kimball Ave	8/15/2006	15600	-87.714036	(41.992042, -87.714036)	North Bound: 7500 / South Bound: 8100
960	5116 North	Ashland Ave	8/22/2006	26700	-87.669587	(41.974889, -87.669587)	North Bound: 13100 / South Bound: 13600
85	6416 South	State St	5/2/2006	19300	-87.62526	(41.777072, -87.62526)	North Bound: 8800 / South Bound: 10500
1116	435 North	La Salle St	9/21/2006	32300	-87.632548	(41.890186, -87.632548)	North Bound: 17700 / South Bound: 14600
871	6891 West	Diversey Ave	8/16/2006	18600	-87.797915	(41.930851, -87.797915)	East Bound: 8600 / West Bound: 8000
674	4034 North	Ashland Ave	8/29/2006	33400	-87.669076	(41.955361, -87.669076)	North Bound: 15700 / South Bound: 17700

Figure 01: Average Daily Traffic Counts Row Data

Data Structure & Quality: These two data sets are in CSV format and also the data sets have tabular structure. In these data sets the information's are represented in a single row header data. In the Traffic data set the data are represented according to daily basis of that year and same for the Weather data as well for the year 2006.

● **Consistency**: The data set used in this project are in a standardizing format which can be used for Consistent datasets and it can facilitate accurate comparisons analytical outcomes.

● **Relevancy**: Both of these data sets are relevance to one another, representing the effects and impact on one another.

● **Validity checks**: The presented data sets has been sanitized properly removing all the null field, empty columns and duplicate values.

● **Accuracy**: The presented data sets are provided from authentic source. Proving us a genuine data sets for that year Traffic and weather conditions.

Data sources licenses & obligations: Traffic and Weather datasets have been collected from public sources and are available for free use.

The City of Chicago dataset is covered under an open data license, it is permitted to use non-commercially for a user. Meteostat 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://dev.meteostat.net/terms.html#use-of-services

Accident Datasets License Terms: https://resources.data.gov/open-licenses/

I shall thus refrain from using the data for any commercial reasons in order to comply with the data responsibilities, making sure that all usage stays within the parameters of non-commercial activity.

1.3 Data Pipeline

Technology: I have used **Python** with the libraries **Numpy**, **Opendatasets**, **Pandas**, and to store the data **SQLite** has been used.

Data transformation steps:

- Both data sets are fetched from their respective URLs using the requests library.
- Traffic datasets is transformed to create a monthly Average Traffic count of the city.
- To study the monthly averages, weather data for the year 2023 is filtered and grouped by month.
- Finally, both data tables are stored in a SQLite database for further analysis.

Table: traffic

	id	month	traffic
	Filter	Filter	Filter
1	1	JAN	104600
2	2	FEB	220400
3	3	MAR	3854900
4	4	APR	1002000
5	5	MAY	698100
6	6	JUN	35200
7	7	JUL	12600
8	8	AUG	6481000
9	9	SEP	3751900
10	10	OCT	6903400
11	11	NOV	1621400
12	12	DEC	211900

Figure 02: Traffic Data Transformed and cleaned

Problems Encountered and Error Handling: Initially, decompressing weather data for retrieval posed challenges. We resolved this by using the **gzip** library to decompress the data before processing. Ensuring data accuracy and consistency was demanding, particularly with missing or incorrect values, but thorough data sanitization and validation resolved these issues.

To track and manage exceptions during data processing, we implemented robust error-handling procedures. The pipeline is designed to adapt dynamically to new data structures or formats, allowing it to handle changing input data smoothly.

1.4 Result and Limitations

Data Output: The data pipeline outputs two tables in a SQLite database: '**weather'**, which contains monthly weather averages, and '**traffic'**, which contains monthly average daily traffic count data for 2006. Both tables are structured with appropriate columns and data types to facilitate effective analysis.

Data Structure and Quality: The accuracy and consistency of the input data are preserved in the output. Simple integration with a variety of frameworks and analysis tools is ensured by using SQLite. This method facilitates effective data retrieval and improves data accessibility. The organized format facilitates easy data reporting and processing.

Limitation and Potential Issues: Even though the data pipeline processes and stores the data successfully, anomalies or outliers may appear during the analysis step. Strict data validation and profiling methods are necessary to find and address these problems. Prior to drawing any conclusions, it is important to verify the integrity of the analysis by using statistical techniques and anomaly detection algorithms.