FAU

Methods of Advanced Data Engineering (MADE)

Data Report

Prepared by Ibne Sayad (23365652)

1.1 Question

What is the impact of **weather** on motor vehicle **accidents** in the city of **Chicago** in 2023?

1.2 Data Sources

I have chosen these data sets to find the impact of weather on road accidents in a particular city Chicago. That's why I have collected weather and road accident data for that city. The analysis will be for the year of **2023**. The weather will be studied based on the seasons after which the impact of it on road accidents will be analyzed.

Data source-01: Weather Data of the City of Chicago

Metadata: https://meteostat.net/de/place/us/chicago?s=72534&t=2023-01-01/2023-12-31

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

Data Type: CSV

This data source contains Chicago's weather report, which has been generated from Meteostst throughout the city from 2023. This data source will provide weather and climate data in Chicago including air temperature, precipitation, snow depth, visibility, wind speed, and additional relevant attributes.

Data source-02: Road Accident Data of the City of Chicago

Metadata: https://catalog.data.gov/dataset/traffic-crashes-vision-zero-chicago-traffic-fatalities

Sample Data: https://data.cityofchicago.org/api/views/gzaz-isa6/rows.csv

Data Type: CSV

This data source contains Chicago's Road Accident details for 2019 to 2024. Among all the years, only 2023 will be considered for analytics. This dataset includes several attributes like Accident Date and Time, Location, Victim, and other relevant attributes.

	A	B	C	D	F	G	H
1	Person_ID	Crash_Date	Crash_Location	Victim	Longitude	Latitude	Location
2	O1538374	03/19/2023 11:30:00 AM	2400 S WENTWORTH AVE	PEDESTRIAN	-87.63207812	41.84911545	POINT (-87.63207812 41.84911545)
3	O699223	07/28/2019 02:13:00 AM	3529 N HALSTED ST	DRIVER	-87.64930836	41.94632714	POINT (-87.64930836 41.94632714)
4	O1506700	01/27/2023 05:46:00 PM	5049 W WASHINGTON BLVD	PEDESTRIAN	-87.75216366	41.88161789	POINT (-87.75216366 41.88161789)
5	O1149843	07/28/2021 07:14:00 PM	7899 S STONY ISLAND	DRIVER	-87.58546514	41.75151786	POINT (-87.58546514 41.75151786)
6	O853602	03-02-20 2:07	700 N LASALLE DR	DRIVER	-87.63282	41.89485	POINT (-87.63282 41.89485)
7	O775752	11-06-19 6:58	3800 N MILWAUKEE AVE	CYCLIST	-87.74195	41.94999	POINT (-87.74195 41.94999)
8	O1290371	02/28/2022 11:18:00 PM	11100 S COTTAGE GROVE	DRIVER	-87.61001554	41.69274215	POINT (-87.61001554 41.69274215)
9	O1131709	07-03-21 1:30	4200 S LAKE SHORE DRIVE	DRIVER	-87.59662067	41.81849736	POINT (-87.59662067 41.81849736)
10	O1114493	06-10-21 11:30	9139 S COMMERCIAL AVE	PEDESTRIAN	-87.55108091	41.7289095	POINT (-87.55108091 41.7289095)
11	O1505719	01/26/2023 12:23:00 PM	335 S LARAMIE AVE	PEDESTRIAN	-87.75474498	41.87571235	POINT (-87.75474498 41.87571235)
12	O989412	11-05-20 23:05	4033 W FULLERTON AVE	MOTORCYCLIST	-87.72816118	41.92447705	POINT (-87.72816118 41.92447705)
13	O1359477	06/16/2022 09:58:00 AM	3351 W 16TH ST	PEDESTRIAN	-87.70990358	41.85892424	POINT (-87.70990358 41.85892424)

Figure 01: Raw accident data from the data source

Data Structure & Quality: Both data sets are tabular structure and in **CSV** format. The accident data includes a single row of header information indicating the information in the column. The weather datasets include weather details based on dates.

● **Accuracy**: The data sets I have chosen are collected from authentic sources. So, these data reflect the real world data, and 100% 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 last year (2023) so that the latest incident can be considered for relevancy.

● **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.

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/

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.

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.
- Accident datasets is transformed to create a monthly total of accidents.
- 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.

	id	month	crashes
	Filter	Filter	Filter
1	1	Jan	12
2	2	Feb	11
3	3	Mar	10
4	4	Apr	8
5	5	May	7
6	6	Jun	11
7	7	Jul	5
8	8	Aug	5
9	9	Sep	11
10	10	Oct	19
11	11	Nov	7
12	12	Dec	9

Figure 01: Transformed accident data from the data source.

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 cleaning 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 holds monthly weather averages, and '**accidents'**, which contains monthly accident data for 2023. Both tables are structured with appropriate columns and data types to facilitate effective querying and analysis.

Data Structure and Quality: The output maintains the consistency and accuracy of the input data. Using SQLite ensures easy integration with various frameworks and analysis tools. This approach also enhances data accessibility and supports efficient data retrieval. The structured format aids in seamless data manipulation and reporting.

Limitation and Potential Issues: In the analysis phase, outliers or anomalies can arise despite the data pipeline's successful processing and storage. Rigorous data profiling and validation techniques are essential to identify and mitigate these issues. Implementing statistical methods and anomaly detection algorithms will help ensure the integrity of the analysis before deriving any insights.