Methods of Advanced Data Engineering (MADE)

Data Report

By

Subroto Karmokar- 23361280

1.1 Question

What insights can be gained from analyzing the correlation between Chicago library visitor data and Chicago weather data for the year 2018?

1.2 Data Sources

Weather Data

- **Data Source 1:** Weather Data of the City of Chicago.
- **Metadata URL:** https://catalog.data.gov/dataset/libraries-2018-visitors-by-location
- **Sample Data:** https://data.cityofchicago.org/api/views/i7zz-iiza/rows.csv
- **Reason for Choice:** The dataset offers weather data for Chicago city.
- **Content:** Weather data for Chicago, including average temperature, precipitation, snowfall, and wind speed, for the year 2018.

Library Visitor Data

- **Data Source 2ss:** Library visitor data of Chicago city.
- **Metadata URL:** https://catalog.data.gov/dataset/libraries-2018-visitors-by-location
- **Sample Data:** https://data.cityofchicago.org/api/views/i7zz-iiza/rows.csv
- **Reason for Choice:** The dataset offers comprehensive monthly visitor counts for Chicago libraries, which can provide information about trends and patterns in library usage.
- **Content:** Chicago library branches monthly visitor counts from January to December 2018.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	LOCATION	JANUARY	FEBRUARY	MARCH	APRIL	MAY	JUNE	JULY	AUGUST	SEPTEMBER	OCTOBER	NOVEMBER	DECEMBER	YTD
2	Albany Park	15687	10569	13383	12505	12000	12966	12637	12890	12309	14294	11525	10795	151560
3	Altgeld	2734	2575	2847	2660	2653	3295	2791	2840	2656	2928	2380	2404	32763
4	Archer Heights*	7060	6469	9119	7224	7091	7566	7774	7905	8624	8514	7341	6807	91494
5	Austin	0	2581	5958	5873	5625	6599	8925	8463	5100	7087	5351	5921	67483
6	Austin-Irving	8619	7541	9578	8469	8243	8703	9800	9477	8403	9331	8804	7702	104670
7	Avalon*	9937	8694	9782	10112	8814	9071	10346	10975	9772	9693	8701	8544	114441
8	Back of the Yards	4342	3443	5152	3920	3834	4008	4767	5208	4295	4765	3976	3297	51007
9	Beverly*	9548	7675	7906	7250	6849	7477	8113	7611	7252	8329	8949	7296	94255
10	Bezazian	9138	8071	10465	9943	10005	11314	12128	12526	9557	10422	9785	9726	123080

Figure 01: Library visitor data (Raw)

Data Structure and Quality

Library visitor dataset is in tabular structure with columns for library location inside Chicago and visitor in all months. On the other hand, weather dataset is also in tabular structure with hourly weather reports. Both datasets are CSV format.

- **Accuracy:** These datasets are collected from authentic sources. Which is reflect the real world data and accurate.
- **Completeness:** All dataset information is complete and very few missing values.
- **Consistency:** Same data format follows in every places.
- **Validity checks:** I have checked and I dropped duplicate data and missing values.

Data Source Licenses

- Both datasets are acquired from public sources and are available for free use.
- The City of Chicago dataset is covered under an open data license, enabling reuse and redistribution for different uses. Meteostat provides weather data licensed under non-commercial terms (CC BY-NC 4.0), allowing for sharing and non-commercial use with proper credit.
- Weather Data License Link: https://dev.meteostat.net/terms.html#use-of-services
- Library Visitor Data License Link: https://resources.data.gov/open-licenses/

1.3 Data Pipeline

Technology Used

- The data pipeline is implemented in Python using libraries like SQLite for data storage and Pandas for data manipulation.

Transformation and Cleaning Steps

- Using the requests library, both data sources are fetched from their respective URLs.
- Data from libraries is transformed to create a monthly total of visitors.
- To compute monthly averages, weather data for the year 2018 is filtered and grouped by month.
- After that, both datasets are stored in a SQLite database for later analysis.

Problems Encountered and Solutions

- Initially, handling compressed data for weather data retrieval presented challenges. This was fixed by first decompressing the data using the gzip library before processing.
- It took a lot of work to ensure data accuracy and consistency, especially when handling missing or incorrect values. Meticulous data cleaning and validation procedures resolved these.

Error Handling and Input Data Changes

- In order to identify and track any exceptions that might happen while processing data, error-handling procedures are implemented.
- The pipeline is designed to dynamically adjust to new data structures or formats in order to gracefully handle changing input data.

	id	month	visitors
	Filter	Filter	Filter
1	1	Jan	628850
2	2	Feb	588219
3	3	Mar	792718
4	4	Apr	715800
5	5	May	681893
6	6	Jun	719255
7	7	Jul	752548
8	8	Aug	794686
9	9	Sep	700404
10	10	Oct	804583
11	11	Nov	709926
12	12	Dec	572282

Figure 02: Library visitor data (Cleaned)

1.4 Result and Limitations

Output Data

- Two tables in the SQLite database represent the output of the data pipeline: weather, which contains monthly weather averages and library, which contains monthly visitor data for the year of 2018.
- To allow effective querying and analysis, both tables have been structured with the right columns and data types.

Data Structure and Quality

- Ensuring consistency and accuracy in the stored information, the outcome preserves the structure and quality of the input data.
- The output format selection of SQLite makes it simple to integrate with a variety of frameworks and analysis tools.

Reflection and Potential Issues

- In the analysis phase, potential problems like outliers or anomalies in the data could arise even though the data pipeline processes and stores the data successfully.
- It's essential to conduct thorough data exploration and validation to identify and address any such issues before drawing conclusions from the analysis.