SOMAIYA VIDYAVIHAR
K J Somaiya Institute of Engineering & Information Technology

Department of Artificial Intelligence &
Data Science Engineering

Academic Year 2022-23

MEDISTATS

PHARMACEUTICAL DRUG SALES ANALYTICS DASHBOARD

Guided by : Prof. Sarika Mane

Group No: 13

Ruchira Patil (Roll No.: 35)

K. J. SOMAIYA INSTITUTE OF ENGINEERING & INFORMATION TECHNOLOGY SION (E), MUMBAI 400022

UNIVERSITY OF MUMBAI

CERTIFICATE

This is to certify that the project titled **MEDISTATS - PHARMACEUTICAL DRUG SALES ANALYTICS DASHBOARD** is completed under my supervision and guidance in partial fulfillment of the requirements of the course AIPR53 Project Based Learning - Mini PR Lab-I, by the following student:

Ruchira Patil (Roll No.: 35)

The course is a part of semester V of the Department of Artificial Intelligence and Data Science during the academic year 2022-2023. The said work has been assessed and is found to be satisfactory.

(Internal guide name and sign.) (External Examiner name and sign.)

College seal

INDEX

1. INTRODUCTION

1.1 Background

The pharmaceutical industry is one of the most data-intensive sectors, generating vast amounts of sales data across hospitals, pharmacies, distributors, and centralised fulfilment systems. Each transaction contributes to datasets that vary by time granularity, geographic scope, and product category. These datasets are essential for understanding market demand, planning inventory, evaluating product performance, and supporting strategic decision-making.

However, raw pharmaceutical sales data is rarely analysis-ready. It often contains inconsistent date formats, duplicated records, missing values, and fragmented structures spread across multiple files or systems. Analysts frequently rely on spreadsheets to analyze this data, but spreadsheets struggle to handle time-series analysis at scale, especially when multiple temporal resolutions are involved.

1.2 Problem Statement

The core problem addressed in this project is the lack of a unified, structured, and interactive system to analyze pharmaceutical sales data across multiple time granularities. Specifically:

- Raw sales data cannot be directly used for analytics due to formatting and consistency issues.
- Spreadsheet-based analysis is error-prone and difficult to scale.
- There is no single view that combines hourly, daily, weekly, and monthly insights.
- Stakeholders lack an intuitive way to explore trends, patterns, and category-wise performance.

1.3 Motivation

The motivation behind this project is to demonstrate how modern analytics tools can transform raw operational data into actionable business insights. By integrating Python-based data preprocessing, relational data modeling principles, and interactive visualization, the project aims to showcase a practical, industry-aligned analytics workflow applicable to real-world pharmaceutical datasets.

1.4 Objectives

- To preprocess and clean raw pharmaceutical sales datasets using Python
- To standardize date formats and remove inconsistencies across datasets
- To design a logical, SQL-style relational data model
- To develop an interactive Power BI dashboard for multi-level sales analysis
- To extract meaningful insights related to sales trends and drug category performance

2. LITERATURE SURVEY

2.1 Pharmaceutical Sales Analytics

Previous research in pharmaceutical analytics highlights the importance of analysing sales data across different time horizons to identify seasonal patterns, demand surges, and long-term trends. Studies emphasise that aggregating sales data at only one level (for example, monthly) can mask important short-term behaviours visible at daily or hourly levels.

2.2 Importance of Time-Series Granularity

Time-series analysis literature suggests that different business questions require different temporal resolutions. Hourly data is useful for understanding peak demand periods, daily data helps identify short-term fluctuations, weekly data smooths volatility, and monthly data supports long-term planning. Effective analytics systems, therefore, require the ability to work with multiple granularities simultaneously.

2.3 Role of Data Cleaning in Analytics

Data quality is a recurring theme in analytics research. Multiple studies report that data preprocessing consumes the majority of analytics effort. Common issues include invalid dates, inconsistent formats, duplicate records, and missing values. Python libraries such as Pandas and NumPy are widely adopted due to their ability to handle structured data transformations programmatically and reproducibly.

2.4 Business Intelligence and Dashboarding

Business intelligence tools like Power BI have been shown to improve decision-making by enabling interactive data exploration. Unlike static reports, dashboards allow users to filter, drill down, and compare metrics dynamically. Research indicates that visual analytics improves comprehension, especially for non-technical stakeholders.

2.5 Relevance to This Project

This project builds upon these concepts by combining data cleaning, time-series aggregation, relational modelling, and visualisation into a single analytics workflow tailored for pharmaceutical sales data.

3. DATA COLLECTION AND PREPROCESSING

This project uses four raw CSV files, each representing a different time granularity:

1 salesdaily.csv
2 salesweekly.csv
3 salesmonthly.csv
4 saleshourly.csv

These files were sourced as raw transactional datasets without prior preprocessing.

3.1 Tools Used

- **Jupyter Notebook**: For interactive data exploration
- **Pandas**: For data manipulation, date parsing, and deduplication.
- **NumPy**: For numerical operations and derived calculations.
- **Python**: For scripting and reproducibility

3.2 Common Data Issues Identified

Across all four files, the following issues were observed:

- Incorrect date formats (MM-DD-YYYY, DD-MM-YYYY, mixed)
- Date columns stored as text
- Duplicate rows
- Error rows caused by invalid dates
- Inconsistent column naming (datum, Date)
- Missing values

3.3 Cleaning Steps – Sales Daily Dataset

The salesdaily.csv file included a column named datum formatted as text in MM/DD/YYYY.

Steps performed:

1. Converted datum column to string
2. Split date into month, day, year
3. Reconstructed a valid datetime column using Pandas
4. Removed rows where date conversion failed
5. Dropped intermediate helper columns
6. Removed duplicate rows
7. Reformatted final date as DD/MM/YYYY
8. Reordered columns to place date first

Power BI performs best when dates are in a recognized date format. Incorrect parsing causes slicers and time intelligence to fail.

	date	m01ab	m01ae	n02ba	n02be	n05b	n05c	r03	r06	year	month	hour	weekday name
0	2/1/2014	0.0	3.67	3.4	32.40	7.0	0.0	0.0	2.0	2014	1	248	Thursday
1	3/1/2014	8.0	4.00	4.4	50.60	16.0	0.0	20.0	4.0	2014	1	276	Friday
2	4/1/2014	2.0	1.00	6.5	61.85	10.0	0.0	9.0	1.0	2014	1	276	Saturday
3	5/1/2014	4.0	3.00	7.0	41.10	8.0	0.0	3.0	0.0	2014	1	276	Sunday
4	6/1/2014	5.0	1.00	4.5	21.70	16.0	2.0	6.0	2.0	2014	1	276	Monday

```
Initial rows: 2106, Cleaned rows: 2106
```

3.4 Cleaning Steps – Sales Weekly Dataset

Issues found:

- Date column stored as text
- Incorrect delimiter interpretation

Fixes applied:

- Applied the same date-splitting logic as daily sales
- Standardized column name to date
- Removed duplicates and errors

	date	M01AB	M01AE	N02BA	N02BE	N05B	N05C	R03	R06
0	5/1/2014	14.00	11.67	21.3	185.95	41.0	0.0	32.0	7.0
1	12/1/2014	29.33	12.68	37.9	190.70	88.0	5.0	21.0	7.2
2	19/1/2014	30.67	26.34	45.9	218.40	80.0	8.0	29.0	12.0
3	26/1/2014	34.00	32.37	31.5	179.60	80.0	8.0	23.0	10.0
4	2/2/2014	31.02	23.35	20.7	159.88	84.0	12.0	29.0	12.0

```
Initial rows: 302, Cleaned rows: 302
```

3.5 Cleaning Steps – Sales Monthly Dataset

This dataset had fewer errors but still required:

- Date type correction
- Removal of duplicate records
- Consistent column naming

	date	M01AB	M01AE	N02BA	N02BE	N05B	N05C	R03	R06
0	31/1/2014	127.69	99.090	152.100	878.030	354.0	50.0	112.0	48.2
1	28/2/2014	133.32	126.050	177.000	1001.900	347.0	31.0	122.0	36.2
2	31/3/2014	137.44	92.950	147.655	779.275	232.0	20.0	112.0	85.4
3	30/4/2014	113.10	89.475	130.900	698.500	209.0	18.0	97.0	73.7
4	31/5/2014	101.79	119.933	132.100	628.780	270.0	23.0	107.0	123.7

```
Initial rows: 70, Cleaned rows: 70
```

3.6 Cleaning Steps – Sales Hourly Dataset

Issues Found:

- datum column causing parsing errors
- Hour column present but date invalid

Resolution:

- Removed rows with invalid dates
- Preserved hour-level granularity
- Standardized column names

	date	Hour	M01AB	M01AE	N02BA	N02BE	N05B	N05C	R03	R06	Year	Month	Weekday Name
0	1/2/2014	8	0.0	0.67	0.4	2.0	0.0	0.0	0.0	1.0	2014	1	Thursday
1	1/2/2014	9	0.0	0.00	1.0	0.0	2.0	0.0	0.0	0.0	2014	1	Thursday
2	1/2/2014	10	0.0	0.00	0.0	3.0	2.0	0.0	0.0	0.0	2014	1	Thursday
3	1/2/2014	11	0.0	0.00	0.0	2.0	1.0	0.0	0.0	0.0	2014	1	Thursday
4	1/2/2014	12	0.0	2.00	0.0	5.0	2.0	0.0	0.0	0.0	2014	1	Thursday

```
Initial rows: 20028, Cleaned rows: 20028
```

3.7 Outcome of Data Cleaning

After completing the data cleaning and preprocessing steps, all four datasets (daily, weekly, monthly, and hourly) were standardized and prepared for analysis. The cleaned datasets:

- Contain valid and consistently formatted date fields
- Are free from duplicate and invalid records
- Share a uniform schema across all time granularities
- Are suitable for aggregation, modeling, and visualization

These outcomes ensured that the datasets were analytics-ready and could be reliably used for downstream data modeling and dashboard development.

```
Data Cleaning Summary:
```

	Dataset	Raw Row Count	Cleaned Row Count	Retention %
0	Daily	2106	2106	100.0
1	Weekly	302	302	100.0
2	Hourly	20028	20028	100.0
3	Monthly	70	70	100.0

```
Cleaned files saved successfully.
```

4. DATA MODELING USING SQL CONCEPTS

4.1 SQL-Style Modeling Approach

Although this project does not use a physical SQL database, it still adopts SQL modelling principles. Each dataset represents a **fact table** at a specific time granularity:

- Sales Hourly → Hour-level facts
- Sales Daily → Day-level facts
- Sales Weekly → Week-level facts
- Sales Monthly → Month-level facts

All tables share common date attributes, which act as logical dimensions.

4.2 Relationship Design

Relationships were created using the date field to enable:

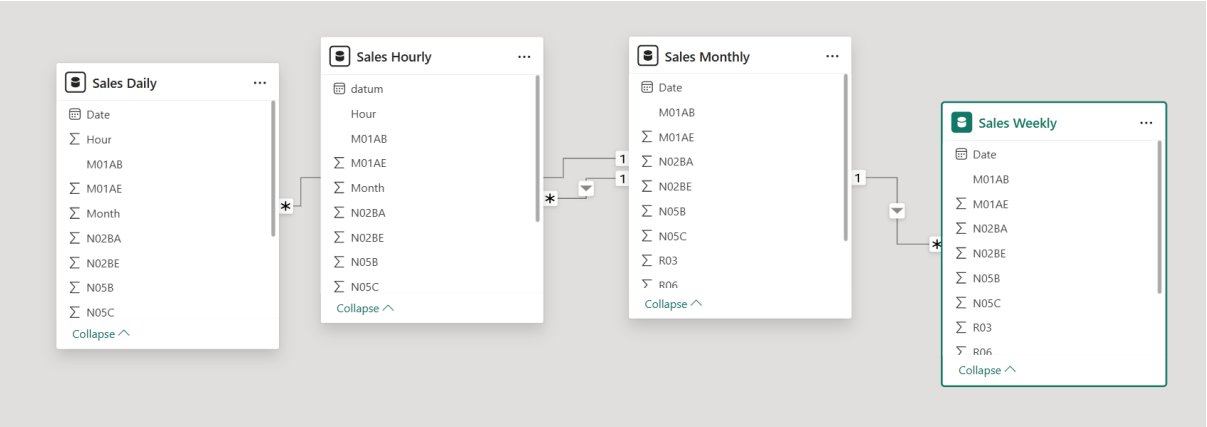
- Cross-filtering between datasets
- Consistent time-based slicing
- Aggregation behavior similar to SQL JOIN operations

This design allows Power BI to compute metrics such as total sales and averages across multiple tables without duplicating data.

4.3 Analytical Benefits

The SQL-style model enables:

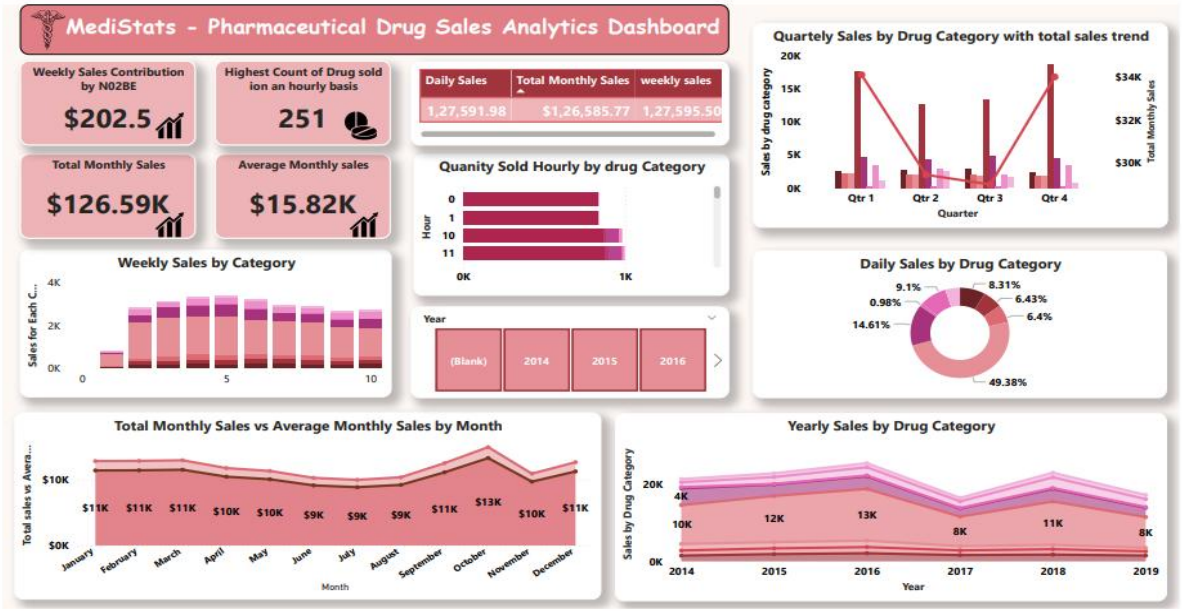
- Efficient aggregation similar to GROUP BY queries
- Clean separation of concerns between data preparation and visualization
- Scalability if the data is later migrated to a relational database



The final model ensures that all visuals respond consistently to filters and slicers, enabling accurate and reliable analysis across time dimensions.

5. DATA ANALYSIS AND VISUALIZATION USING POWER BI

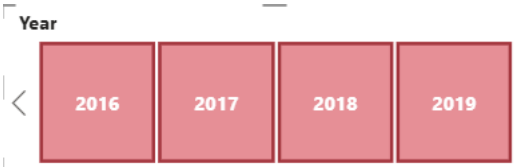
The MediStats Power BI dashboard was designed to present pharmaceutical sales data in a structured and interactive manner, enabling analysis across multiple time granularities and drug categories. The dashboard integrates summary-level indicators, comparative charts, and time-based visualizations, all driven by explicitly defined calculations to ensure accuracy and consistency.



This dashboard consolidates daily, weekly, monthly, and hourly pharmaceutical sales data into a single interactive analytics interface. The visuals are driven by DAX measures and slicers that allow users to explore sales patterns across time, drug categories, and aggregation levels.

5.1 Year-wise Dashboard Context (2014–2019)

The dashboard includes a year slicer covering the period from **2014 to 2019**, which filters all visuals simultaneously. This allows users to analyze sales performance for a specific year or compare trends across multiple years without modifying the underlying data model.



5.2 Total Monthly Sales

Total Monthly Sales represents the overall sales value aggregated across all drug categories at the monthly level. Since sales for each drug category are stored in separate columns in the dataset, the total value is calculated by explicitly summing the sales of all eight drug categories.

Formula used: Total Monthly Sales = SUM of all 8 medicine

This calculation ensures that monthly sales values are aggregated transparently and consistently across all drug categories.

Daily Sales	Total Monthly Sales	weekly sales
1,27,591.98	$1,26,585.77	1,27,595.50

5.3 Average Monthly Sales

Average Monthly Sales provides a normalized view of monthly sales by calculating the average contribution per drug category. This metric is derived by dividing the total monthly sales value by the number of drug categories present in the dataset.

Formula used:
Average Monthly Sales = DIVIDE([Total Monthly Sales], 8, 0)

Here, the value **8** represents the total number of drug categories. The DIVIDE function is used to safely handle potential divide-by-zero cases.

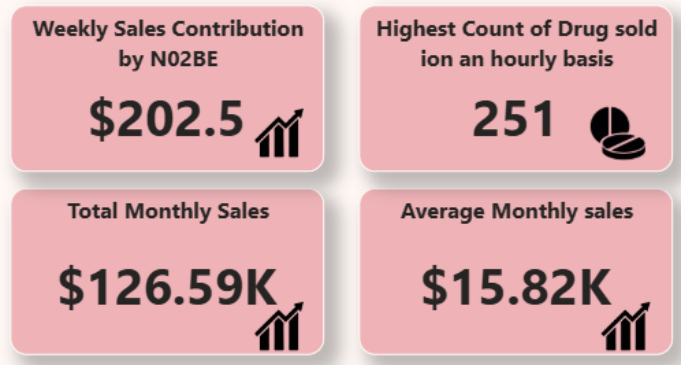
5.4 Daily Sales

Daily Sales represents the aggregated sales value at the daily level. This metric is computed by summing daily sales values across all drug categories from the Sales Daily dataset. It provides visibility into short-term sales activity and day-level demand patterns.

5.5 Weekly Sales

Weekly Sales represents aggregated sales values at the weekly level, combining sales across all drug categories. This calculation smooths daily-level volatility while still preserving meaningful temporal variation.

Formula used:
weekly sales = SUM('Sales Weekly'[M01AB]) + SUM('Sales Weekly'[M01AE]) + SUM('Sales Weekly'[N02BA]) + SUM('Sales Weekly'[N02BE]) + SUM('Sales Weekly'[N05B]) + SUM('Sales Weekly'[N05C]) + SUM('Sales Weekly'[R03]) + SUM('Sales Weekly'[R06])



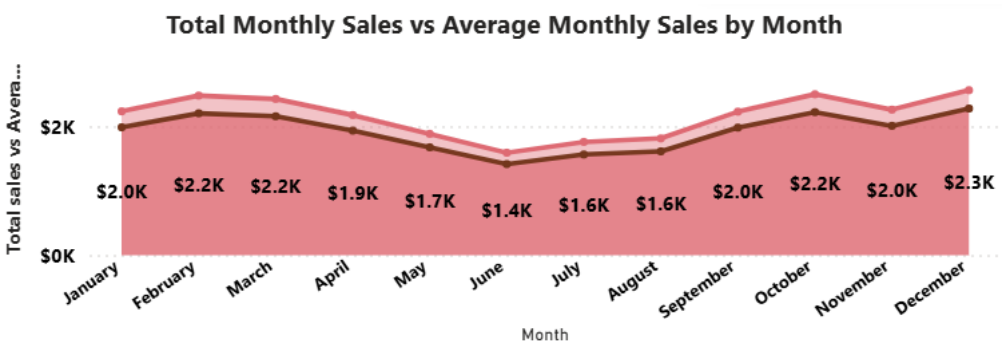
5.6 Weekly Sales Contribution by Drug Category (N02BE)

The dashboard includes a metric that calculates the percentage contribution of the **N02BE** drug category to total weekly sales. This is achieved by dividing weekly sales by the sales of the N02BE category and converting the result into a percentage.

Formula used:
weekly sales contribution by N02BE = DIVIDE('Sales Weekly'[weekly sales],SUM('Sales Weekly'[N02BE]),0) * 100

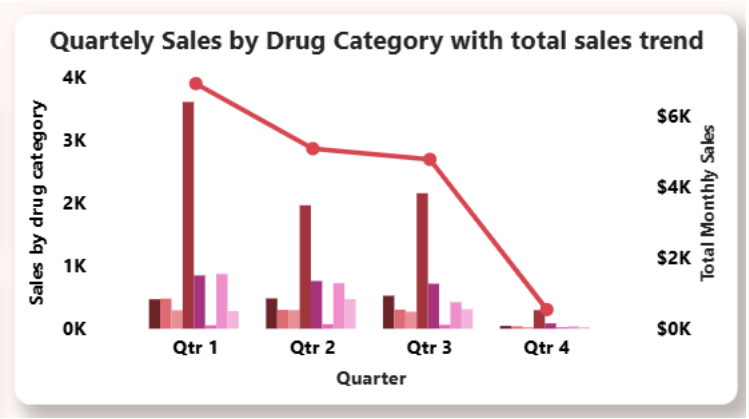
5.7 Total Monthly Sales vs Average Monthly Sales by Month

This visualization compares total monthly sales against the calculated average monthly sales across all months in 2018. A stacked area chart with an overlaid line is used to display how actual sales fluctuate relative to the average baseline over time.



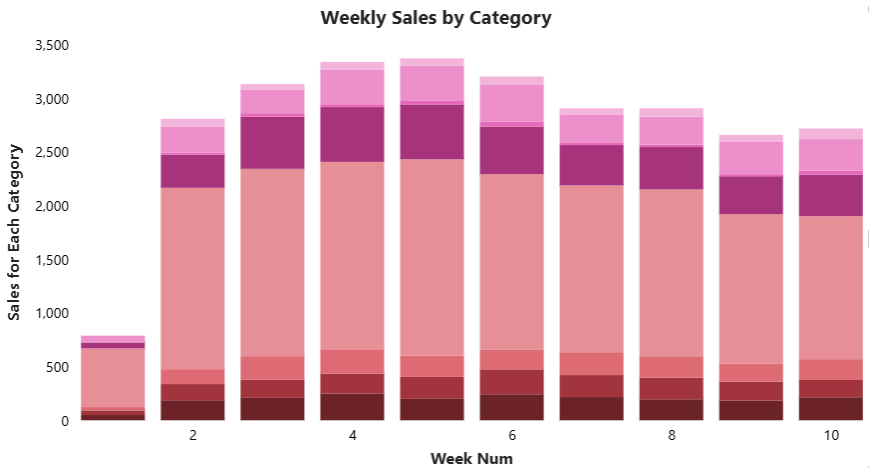
5.8 Quarterly Sales by Drug Category with Total Sales Trend

Quarterly sales are visualized using a clustered column chart, with an overlaid line representing total quarterly sales. This combination allows comparison of individual drug category performance while also highlighting overall sales trends across quarters in 2019.



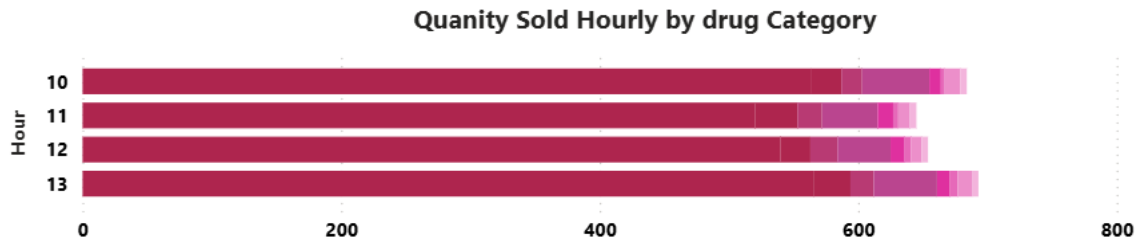
5.9 Weekly Sales by Drug Category

Weekly sales distribution across drug categories is visualized using a stacked column chart. This chart shows both total weekly sales and the relative contribution of each drug category within a given week.



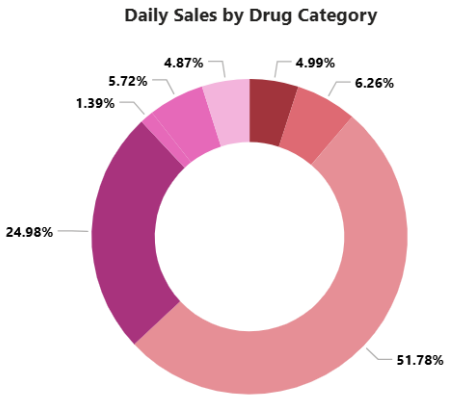
5.10 Quantity Sold Hourly by Drug Category

Hourly sales patterns are visualized using a horizontal bar chart, which displays the quantity of drugs sold at different hours of the day. This representation enables easy comparison of sales volume across time intervals.



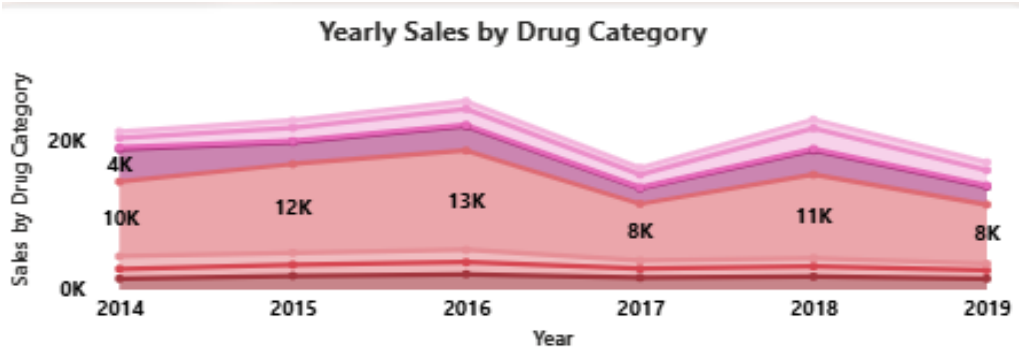
5.11 Daily Sales by Drug Category

Daily sales distribution is represented using a donut chart, showing the percentage contribution of each drug category to total daily sales. This provides a proportional view of category-wise performance at the daily level in 2017.



5.12 Yearly Sales by Drug Category (2014–2019)

Yearly sales trends across all drug categories are visualized using a multi-line chart spanning the years 2014 to 2019. This chart provides a long-term view of sales performance and highlights changes in category behavior over time.



6. CONCLUSION

This project successfully demonstrates an end-to-end pharmaceutical sales analytics workflow, transforming raw transactional data into meaningful, decision-ready insights using modern data analytics and visualization tools. Four heterogeneous datasets—daily, weekly, monthly, and hourly sales—were systematically cleaned, standardized, and integrated to ensure consistency across time granularities. Python-based preprocessing using Pandas and NumPy played a critical role in correcting date formats, removing duplicates, handling missing values, and aligning schemas, resulting in datasets that were reliable and analytics-ready.

A centralized data model was designed to support time-series analysis and aggregation across multiple dimensions, including drug categories, time periods, and sales frequency. The introduction of a unified Date dimension enabled consistent filtering and slicing across all datasets, resolving granularity mismatches and improving analytical accuracy. Key performance indicators such as Total Monthly Sales, Average Monthly Sales, Daily Sales, and Weekly Contribution were calculated using DAX expressions to ensure clarity, traceability, and scalability of metrics.

The Power BI dashboard developed as part of this project provides an interactive and intuitive interface for exploring pharmaceutical sales trends. Multiple visualization types—including KPI cards, stacked column charts, area charts, donut charts, and combined column-line charts were carefully selected to highlight different analytical perspectives such as seasonality, category dominance, hourly demand patterns, and long-term growth trends. By replacing spreadsheet-based analysis with a centralized dashboard, the solution significantly improves interpretability, reduces manual effort, and enables faster insight generation.

Overall, the project validates the effectiveness of combining Python-based data preprocessing, structured data modeling, and interactive visualization to analyze complex, real-world healthcare sales data. The approach adopted in this study is scalable, reproducible, and applicable to broader business intelligence and healthcare analytics use cases.

7. LIMITATIONS AND FUTURE SCOPE

7.1 Limitations

1. **Synthetic Dataset Constraint**
   The dataset used in this project is publicly available and simulated, which may not fully capture real-world complexities such as pricing variability, prescription policies, or supply-chain disruptions.

2. **Limited Business Context**
   The analysis focuses primarily on sales quantities and values, without incorporating external factors such as marketing campaigns, demographic data, or regulatory changes that influence pharmaceutical demand.

3. **Hourly Data Quality Issues**
   A significant proportion of duplicate records in the hourly dataset suggests potential logging or data collection inconsistencies, limiting the reliability of fine-grained temporal analysis.

4. **Static Dashboard Outputs**
   While the dashboard is interactive, it does not currently support real-time data ingestion or live updates from transactional systems.

7.2 Future Scope

1. **Integration of Real-Time Data Pipelines**
   The dashboard can be extended to ingest live pharmacy or hospital data streams using APIs or cloud-based ETL tools for near real-time analytics.

2. **Advanced Forecasting and Predictive Analytics**
   Machine learning models can be incorporated to forecast drug demand, identify seasonal trends, and predict stock-out risks.

3. **Geographical and Demographic Analysis**
   Adding region-level, city-level, or patient demographic data would enable deeper insights into consumption patterns and healthcare accessibility.

4. **Role-Based Dashboard**
   Separate dashboards can be designed for pharmacists, supply-chain managers, and healthcare administrators, each tailored to specific decision-making needs.

5. **Enhanced Data Governance**
   Implementing data validation rules, anomaly detection, and audit logging would improve data reliability and compliance in production environments.

8. REFERENCES

[1] Frestel, J., Teoh, S.W.K., Broderick, C., Dao, A. and Sajogo, M., 2023. A health integrated platform for pharmacy clinical intervention data management and intelligent visual analytics and reporting. *Exploratory Research in Clinical and Social Pharmacy*, *12*, p.100332.

[2] McKinney, W., 2012. *Python for data analysis: Data wrangling with Pandas, NumPy, and IPython*. " O'Reilly Media, Inc.".

[3] Provost, F. and Fawcett, T., 2013. *Data Science for Business: What you need to know about data mining and data-analytic thinking*. " O'Reilly Media, Inc.".

[4] Kimball, R. and Ross, M. (2013) *The Data Warehouse Toolkit: The Definitive Guide to Dimensional Modeling*. 3rd edn. Hoboken: Wiley.

[5] Belghith, M., Ammar, H.B., Elloumi, A. and Hachicha, W., 2024, May. A new rolling forecasting framework using Microsoft Power BI for data visualization: A case study in a pharmaceutical industry. In *Annales Pharmaceutiques Françaises* (Vol. 82, No. 3, pp. 493-506). Elsevier Masson.

[6] Pianalytix (2023) *Public pharmaceutical sales dataset and Power BI dashboard reference implementation*.

[7] Atobatele, O.K., Ajayi, O.O., Hungbo, A.Q. and Adeyemi, C., 2022. Improving strategic health decision-making with SQL-driven dashboards and Power BI visualization models. *Shodhshauryam Int Sci Refereed Res J*, *5*(5), pp.291-313.

[8] Han, J., Kamber, M. and Pei, J. (2011) *Data Mining: Concepts and Techniques*. 3rd edn. San Francisco: Morgan Kaufmann.

[9] Vest, M.H., Colmenares, E.W. and Pappas, A.L., 2021. Transforming data into insight: establishment of a pharmacy analytics and outcomes team. *American Journal of Health-System Pharmacy*, *78*(1), pp.65-73.