INTEGRATED INTERNSHIP - Final Report

Company	**Yadgreen (*Aramco Namaat Partner*)** focus on Sustainable agriculture, CO_2/GHG reduction and carbon plantations (forestry & mangroves)
Project	Analysing real-world Traffic Data
Tool	Power BI Desktop
Duration	8 Weeks

Case Study Report

CONTENTS

Case Study Report

1. About The Internship & Project Context

This case study represents an 8-week Integrated Data Science Internship, focused on analyzing real-world traffic data using Power BI to support smart city traffic management.

The dataset provided contains valuable traffic information that can be used for analysis and forecasting.

The project followed a structured workflow, including:

- *Data exploration*
- *Data cleaning*
- *Data modeling*
- *Exploratory Data Analysis (EDA)*
- *Data visualization*
- *Final dashboard development and publishing.*

This provided hands-on experience in data analysis, DAX calculations, interactive visualization, prediction, and traffic trend analysis using a real-world traffic volume dataset.

By the end of the study, an interactive Power BI dashboard was developed to showcase actionable insights for improving traffic flow, congestion management, and forecasting future trends.

2. Problem Statement

Traffic congestion has become a major challenge due to increasing vehicle volume and inefficient traffic management, leading to longer travel times, higher fuel consumption, and increased environmental impact from vehicle emissions.

Although traffic data is available, it is not effectively analyzed or visualized, making it difficult to identify congestion patterns, peak traffic hours, and high-traffic junctions for informed decision-making.

3. Project Objective

The main objective of this project is to design a structured data model and an interactive Power BI dashboard that transforms raw traffic data into meaningful insights to:

- Analyze traffic flow over time
- Identify peak traffic hours
- Detect congestion hotspots
- Forecast future traffic trends

Case Study Report

4. Approach & Methodology

To achieve the project objectives, the following approach was adopted:

4.1 Data Preparation

From the given dataset, raw traffic data was imported into Power BI Desktop. Initial data quality checks were performed to identify missing values, duplicates, and data type inconsistencies

- Column quality and profiling confirmed 100% valid data with no missing or null values (column include datetime, ID, Vehicle count and Junction).
- Date and time fields were standardized into a uniform format (YYYY-MM-DD HH:MM:SS) to ensure consistent filtering and time-based analysis.
- Additional derived columns such as Hour, Day, Month, Weekday/Weekend, and Traffic Category (Low, Medium, High) were created to support detailed analysis.
- Outliers were statistically identified using the IQR method and flagged for analysis, while retaining all original values to preserve real-world traffic behavior.

4.2 Data Modeling

A star schema data model was designed to ensure efficient analysis and scalability.

- The traffic dataset was treated as the **fact table**, containing transactional traffic volume records.
- A **Date dimension table (Dim_Date)** was created using DAX to support time-based hierarchies.
- A one-to-many relationship was established between the Date dimension and the fact table using a surrogate DateKey.
- Time hierarchies (Year → Month → Day) were defined to enable drill-down and trend analysis across different time granularities

4.3 DAX Calculations

DAX measures and calculated columns were developed to derive meaningful insights and support advanced analysis.

Calculated Column

```
DateKey = VALUE(FORMAT('traffic dataset'[DateTime], "yyyyMMdd"))

WeekDay_Weekend = IF(WEEKDAY('traffic dataset'[Date],1)=1 || WEEKDAY('traffic
dataset'[Date],1)=7,"Weekend","Weekday")

CategoryOfVehical = SWITCH(TRUE(), [Vehicles] <= 9, "Low", [Vehicles] <= 29, "Medium","High")

OutlierRecord = IF('traffic dataset'[Vehicles] > 59,"Outlier", "Normal")

AdjustedVehicles = IF('traffic dataset'[Vehicles] > 59,59,'traffic dataset'[Vehicles])
```

Case Study Report

```
TimePeriod =
SWITCH(
    TRUE(),
    'traffic dataset'[Hour] = 0, "Midnight",
    'traffic dataset'[Hour] >= 1 && 'traffic dataset'[Hour] <= 3, "Late Night",
    'traffic dataset'[Hour] >= 4 && 'traffic dataset'[Hour] <= 8, "Early Morning",
    'traffic dataset'[Hour] >= 9 && 'traffic dataset'[Hour] <= 11, "Morning",
    'traffic dataset'[Hour] >= 12 && 'traffic dataset'[Hour] <= 16, "Afternoon",
    'traffic dataset'[Hour] >= 17 && 'traffic dataset'[Hour] <= 19, "Evening",
    'traffic dataset'[Hour] >= 20 && 'traffic dataset'[Hour] <= 23, "Night"
)

TimePeriodOrder =
SWITCH(
    TRUE(),
    'traffic dataset'[Hour] = 0, 1,
    'traffic dataset'[Hour] >= 1 && 'traffic dataset'[Hour] <= 3, 2,
    'traffic dataset'[Hour] >= 4 && 'traffic dataset'[Hour] <= 8, 3,
    'traffic dataset'[Hour] >= 9 && 'traffic dataset'[Hour] <= 11, 4,
    'traffic dataset'[Hour] >= 12 && 'traffic dataset'[Hour] <= 16, 5,
    'traffic dataset'[Hour] >= 17 && 'traffic dataset'[Hour] <= 19, 6,
    'traffic dataset'[Hour] >= 20 && 'traffic dataset'[Hour] <= 23, 7
)
```

Calculated Measure

```
Total Vehicles = SUM('traffic dataset'[Vehicles])

2015_Sum = CALCULATE(SUM('traffic dataset'[Vehicles]),'traffic dataset'[Year]=2015)

2016_Sum = CALCULATE(SUM('traffic dataset'[Vehicles]),'traffic dataset'[Year]=2016)

2017_Sum = CALCULATE(SUM('traffic dataset'[Vehicles]),'traffic dataset'[Year]=2017)

25percentile = PERCENTILEX.INC ('traffic dataset', 'traffic dataset'[Vehicles], 0.25)

50percentile = PERCENTILEX.INC ('traffic dataset', 'traffic dataset'[Vehicles], 0.50)

75percentile = PERCENTILEX.INC ('traffic dataset', 'traffic dataset'[Vehicles], 0.75)

IQR = [75percentile] - [25percentile]

LowerBound = [25percentile] - 1.5 * [IQR]

UpperBound = [75percentile] + 1.5 * [IQR]

OutlierCount =
VAR UB = [UpperBound]  -- store the measure in a variable
RETURN
COUNTROWS(FILTER('traffic dataset','traffic dataset'[Vehicles] > UB))

MA_30_Days = AVERAGEX(DATESINPERIOD('traffic dataset'[Date],MAX('traffic dataset'[Date]),-
30,DAY),[Total Vehicles])

Forecast Traffic = [MA_30_Days]

Error = ABS([Total Vehicles] - [Forecast Traffic])

MA_7_Days = AVERAGEX(DATESINPERIOD('traffic dataset'[Date], MAX('traffic dataset'[Date]), -7,
DAY), [Total Vehicles])

MAE = AVERAGEX(VALUES('traffic dataset'[Date]), [Error])
```

4.4 Visualization Design

The dashboard was designed with a focus on clarity, insight delivery, and usability.

- KPI cards were placed at the top to highlight key traffic metrics.
- Line charts, heat maps, pie charts, column chart, and area charts were used to visualize trends, congestion patterns, and category distributions.
- Consistent color themes, labels, and formatting were applied to enhance readability and interpretation.
- Visuals were arranged logically to guide from high-level insights to detailed analysis

4.5 Interactivity Features

Interactive elements were implemented to enable dynamic exploration of traffic data.

- Slicers and filters for Year, Month, Day, Hour, Junction, Weekday/Weekend, and Traffic Category were added.
- Drill-down and drill-through capabilities allowed users to analyze traffic behavior at multiple levels of detail.
- Custom tooltips were used to provide contextual insights without cluttering the dashboard

4.6 Performance Optimization:

To ensure a smooth and responsive user experience, performance optimization techniques were applied.

- Power BI Performance Analyzer was used to identify slow-loading visuals.
- DAX measures were optimized to reduce unnecessary calculations.
- Most visuals achieved load times under acceptable thresholds, ensuring efficient dashboard interaction.

4.7 Validation & Finalization:

The final dashboard was validated for accuracy, usability, and consistency.

- The report was prepared as a ready-to-share Power BI Desktop (.pbix) file, ensuring deployment readiness once Power BI Service access is available.

5. Dashboard Overview

The Smart City Traffic Management dashboard was developed in Power BI to provide a comprehensive and interactive view of traffic patterns across multiple junctions and time periods. The dashboard is designed to support data-driven decision-making by enabling users to quickly identify congestion trends, peak traffic hours, and high-impact junctions.

At the top of the dashboard, KPI cards summarize key traffic metrics such as Distinct traffic records, average vehicles per hour, maximum and minimum vehicle counts, the number of junctions analyzed and total traffic volume in year 2015/2016/2017. These KPIs provide an instant high-level snapshot of overall traffic conditions.

The central section of the dashboard focuses on traffic trends and congestion analysis. Line charts visualize traffic volume over time, yearly, monthly and hourly level. Traffic trends with drill-down capability from year to month and day, allowing users to analyze both long-term and short-term trends. Heat maps display congestion intensity across hours and junctions, clearly highlighting peak traffic periods and recurring congestion patterns.

To analyze traffic composition, category distribution visuals such as pie charts show the proportion of Low, Medium, and High traffic categories. Scatter plots and comparative visuals further help identify unusual traffic spikes and junction-wise variations.

Interactive slicers and filters are integrated throughout the dashboard, enabling users to dynamically explore data by year, month, day, hour, junction, weekday/weekend, and traffic category. Custom tooltips enhance user experience by providing detailed contextual information on hover without overcrowding the visuals.

Overall, the dashboard offers a fully interactive, performance-optimized, and user-friendly analytical interface that enables efficient exploration of traffic data and supports smart city traffic planning, congestion management, and operational decision-making.

6. Key Insights

The interactive dashboard and detailed analysis of traffic data revealed several important insights related to traffic behavior, congestion patterns, and temporal trends:

6.1 Peak Traffic Hours

Traffic volume shows consistent daily peaks during **late morning (10 AM – 12 PM)** and **evening hours (7 PM – 8 PM)** across all junctions. These peaks reflect regular commuting patterns and indicate critical time windows where congestion management measures are most required.

6.2 Junction-Wise Congestion Hotspots

Among all monitored locations, **Junction 1 consistently records the highest traffic volume**, making it the primary congestion hotspot. Junctions 2 and 3 experience moderate traffic levels, while Junction 4 remains relatively low in traffic volume. This highlights Junction 1 as a priority area for traffic control and infrastructure improvements.

6.3 Weekday vs Weekend Traffic Behavior

Traffic volumes are significantly higher on **weekdays** compared to weekends. This pattern confirms that traffic flow is largely driven by work-related and routine travel, whereas weekend traffic is lower and more evenly distributed throughout the day.

6.4 Traffic Category Distribution

The majority of traffic falls under the **medium traffic category**, indicating steady and consistent vehicle flow across most time periods. High traffic categories occur less frequently but are concentrated during peak hours, contributing disproportionately to congestion levels.

6.5 Seasonal and Long-Term Trends

Yearly analysis shows a sharp increase in traffic from **2015 to 2016**, reflecting rising traffic demand. The apparent decline in 2017 is attributed to incomplete data coverage rather than an actual reduction in traffic. Monthly trends reveal higher traffic volumes during **mid-year months**, indicating seasonal variation in traffic flow.

6.6 Outlier and Anomaly Patterns

Outlier detection using the IQR method identified 3617 traffic spikes. **Winsorization was demonstrated** to cap extreme values, but **outliers were not removed** as they represent real-world events such as high demand of utilizing personal vehicles, festivals, accidents, or road closures.

6.7 Forecasting & Prediction Insights

The **7-day moving average** captures short-term traffic fluctuations and weekday–weekend patterns, while the **30-day moving average** smooths short-term noise and highlights the underlying long-term traffic trend.

The moving-average–based forecasting model effectively captures regular traffic behavior; however, **forecast errors increase during sudden traffic spikes** caused by external factors such as events or disruptions. The model achieved a **Mean Absolute Error (MAE) of approximately 267 vehicles**, indicating reasonable accuracy for trend-based forecasting.

7. Proposed Solutions

Based on the insights obtained from the traffic analysis, dashboard observations, outlier assessment, and time-series forecasting, the following solutions are proposed:

- **Optimize traffic signal timings during peak hours** identified in the analysis (10–12 AM and 7–8 PM) to reduce congestion and improve traffic flow.
- **Allocate traffic control resources to high-traffic junctions**, particularly Junction 1, which consistently records the highest traffic volume.
- **Use historical traffic trends and moving-average forecasts (7-day and 30-day)** to support proactive traffic planning, resource allocation, and congestion management.
- **Leverage outlier analysis for event-based traffic planning**, as extreme traffic values represent real-world conditions such as festivals, accidents, or road disruptions and should not be removed from the dataset.
- **Enhance future analysis by integrating external data sources** such as weather conditions, public events, and roadwork schedules to improve forecast accuracy and decision-making.

Case Study Report

8. Publishing, Collaboration & Feedback

The final Power BI dashboard was prepared following best practices for sharing and collaboration. Due to the absence of an organizational Power BI Service account, direct online publishing was not possible. Instead, the finalized dashboard was shared as a **Power BI Desktop (.pbix) file** using a Google Drive link:

https://drive.google.com/file/d/1RL0GpvwY1DycpG1PCye74j6V3PE93w7S/view?usp=drive_link

The report structure, data model, and visual design were created to support easy review and understanding. Sharing the .pbix file helped reviewers explore the dashboard, validate insights, and provide feedback.

9. Limitations

- Traffic data for 2015 and 2017 is partially available, which affects long-term trend interpretation.
- Geospatial analysis was not implemented due to the absence of latitude and longitude information.
- The predictive model is based on historical trends and moving averages, and does not account for sudden external disruptions.
- Lack of Power BI Service account due to this publishing in workspace was not done

10. Conclusion

This project successfully demonstrates the application of data analytics and Power BI for smart city traffic management. Through systematic data preparation, robust data modeling, advanced DAX calculations, and interactive visualization design, meaningful traffic insights were derived.

The dashboard effectively highlights peak traffic hours, congestion hotspots, and long-term traffic trends, while the predictive analysis provides a foundation for proactive traffic planning. Despite data limitations, the solution remains practical, scalable, and ready for deployment.

Overall, this project showcases how data-driven dashboards can support informed decision-making, improve traffic management strategies, and contribute to the development of smarter and more efficient urban transportation systems.

11. Link

PowerBI

PPT