LONDON TRAFFIC EDA REPORT

D Abiodun
12696259

Enterprise Data Analysis Strategy: London Road Traffic Statistics

1. Brief Description of Traffic in the London Region

London, one of Europe's most densely populated cities, faces significant traffic challenges due to high urban density, aging infrastructure, and increased vehicle ownership. The Department for Transport (2023) notes that congestion is worst on major A roads and in Inner London boroughs. The rise in Light Goods Vehicles (LGVs) is driven by e-commerce demands (Bhuyan et al., 2021), with traffic patterns showing higher inbound morning volumes and outbound evening traffic, especially in Westminster, Camden, and Lambeth. Infrastructure development has not kept pace with urban growth, exacerbating these issues (Wang et al., 2009).

2. Detailed Description of a Typical EDA Model Emphasising Storage

This report uses the CRISP-DM model to guide the EDA strategy, backed by a scalable, cloud-native architecture, organised into five key storage layers.

2.1. Data Lake Layer

- Uses platforms such as AWS S3 or Azure Data Lake Storage to preserve raw CSV and Excel datasets.

- Enables schema-on-read flexibility for exploratory querying and archival access (Nambiar & Mundra, 2022).

2.2. ETL and Transformation Layer

- ETL pipelines developed in Python using Pandas and orchestrated via Apache Airflow.

- Incorporates schema validation, null handling, and logging (Coleman et al., 2016).

2.3. Data Warehouse Layer

- Implements schema-on-write with tools such as Azure Synapse.

- Houses a star schema with a central Fact Table (traffic volume by segment and time) and Dimension Tables (borough, road type, vehicle class) (Sawadogo & Darmont, 2021).

2.4. Analytics Layer

- Tableau and Power BI are used for visual analytics, integrated with live warehouse queries.

- Supports spatial (choropleth maps) and temporal (trend plots) representations.

2.5. Governance Layer

- Includes Azure Purview for metadata tracking and access management.

- Ensures data integrity, version control, and compliance (Li et al., 2015).

This architecture allows for horizontal scaling and efficient, compliant traffic data analysis (IDC, 2011).

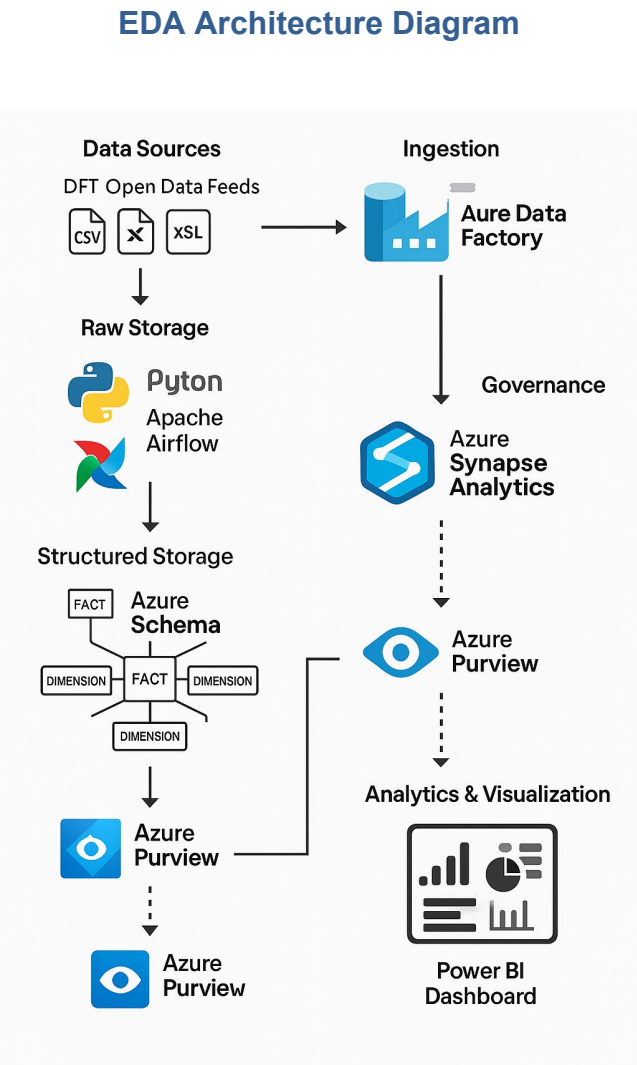

Figure 1: Architecture flowchart for the London Road Traffic EDA Pipeline using Azure components and open-source tools.

3. Step-by-Step Data Analysis Methods Undertaken

The modified CRISP-DM steps included:

1. Business Understanding: Identify high-density congestion and directional traffic imbalance.

2. Data Discovery and Acquisition: Sources include `dft_aadf`, `rawcount`, and `countpoints` (DfT, 2023).

3. Preparation: Data cleaning, date formatting, joining datasets using `count_point_id` and `year`.

4. Exploration: Applied correlation analysis, line graphs, and heatmaps using Python (Gogtay & Thatte, 2017).

5. Evaluation: Results were reviewed with policy relevance in mind (Keller et al., 2020).

Deviation: Advanced filtering and imputation were implemented due to missing road_category_id entries and inconsistent time formats. Group-based imputation for categorical fields and rolling averages for numerical data were used to maintain temporal patterns in sensor-derived time series, following big data analytics standards (Dwivedi, Bali, & James, 2018).

Specifically:

- **Conditional Joins** were used to infer missing road_category_id values by merging related records from the dft_countpoints dataset using count_point_id and road_name.
- **Regex-Based Time Parsing** and pandas.to_datetime() were employed to standardise inconsistent time formats, enabling consistent temporal analysis.
- **Null Threshold Filtering** excluded columns with more than 40% missing data, particularly in direction-specific traffic counts.
- **Group-Based Mode Imputation** was used to fill missing categorical fields, such as road_category_id, based on borough and road name groupings.
- **Rolling Average Imputation** was applied to missing vehicle count values using a 3-day rolling window to smooth irregularities in time series trends.

These techniques provided cleaner model inputs, improving the reliability of downstream analytics in urban traffic. While the CRISP-DM methodology offers a solid foundation, integrating agile, iterative cycles from the OSEMN framework—particularly in modeling and interpretive stages—could enhance adaptive updates, real-time traffic feeds, and continuous improvement while ensuring governance and transparency.

3.1 Methodology Justification and Comparative Evaluation

This project adopts a hybrid methodology combining the Cross-Industry Standard Process for Data Mining (CRISP-DM) and the OSEMN model (Obtain, Scrub, Explore, Model, Interpret). This blend leverages CRISP-DM's structured governance with OSEMN's agility to address the dynamic demands of London's traffic ecosystem. CRISP-DM divides the data lifecycle into phases: Business Understanding, Data Understanding, Preparation, Modelling, Evaluation, and Deployment, supporting auditability and compliance with regulations like UK GDPR. Enterprise tools such as Azure Synapse, Apache Airflow, and Azure Purview enhance its compatibility with scalable, governed data environments.

London’s road network faces variable, real-time conditions like congestion and weather disruptions, necessitating a flexible analytical approach. To tackle this, OSEMN has been integrated into the later stages of the CRISP-DM pipeline, especially during deployment and feedback. Its iterative process enables the rapid ingestion, exploration, and interpretation of sensor inputs, such as ANPR, GPS, and IoT signals. (Dwivedi, Bali, & James, 2018).

While CRISP-DM is optimal for long-term planning and data governance, OSEMN provides operational adaptability. Rather than treating these frameworks as mutually exclusive, this hybrid model leverages their respective strengths: CRISP-DM governs the architecture and modelling logic, while OSEMN enables real-time adjustment, edge-model execution (e.g., TinyML), and feedback loops to inform policy in near real-time.

Criteria	CRISP-DM (Chosen)	OSEMN (Alternative)
Governance & Compliance	Strong alignment with Purview, GDPR	Lacks formal governance phases
Business-Policy Fit	Structured stakeholder alignment	Lightweight, operational-level focus
Real-Time Capability	Limited; built for batch processes	High; ideal for sensor-driven feedback
Tool Compatibility	Azure-native (Synapse, Data Factory)	Open-source & Python friendly (Pandas, Airflow)
Modularity & Reusability	High – ideal for large datasets	High – ideal for fast prototyping

The feedback on the Unit 9 submission highlighted strong design and modeling but called for a deeper justification of the methodology. While the previous report favored CRISP-DM for its structure, this analysis proposes a hybrid approach that combines CRISP-DM's governance with OSEMN's flexibility. This revised method integrates CRISP-DM for foundational development with agile OSEMN iterations for real-time features like IoT feeds and TinyML traffic predictions (Dutta & Bharali, 2021), effectively addressing the complexities of London traffic.

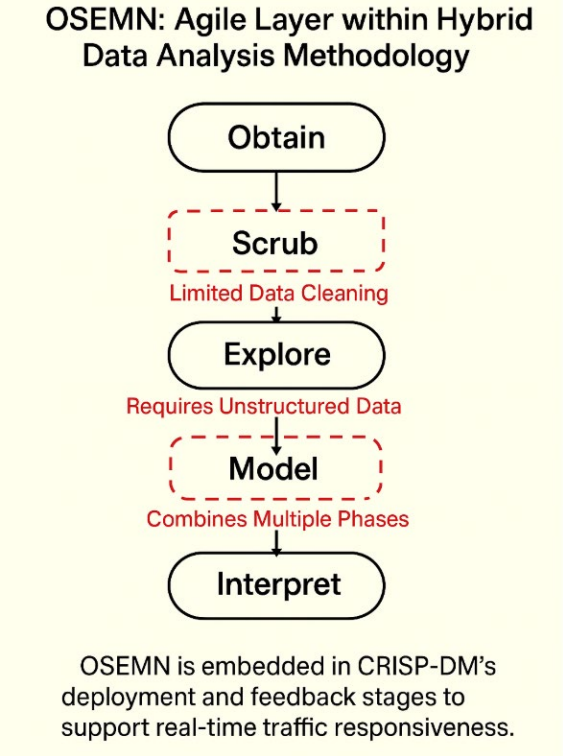

Figure 2: OSEMN Framework

4. Data Analysis Models: Results and Presentation

Model 1: Time Series Trend Analysis

Vehicle count patterns show dual peaks on weekdays (7–9 AM and 4–6 PM).

LGVs showed midday peaks, tied to delivery demand.

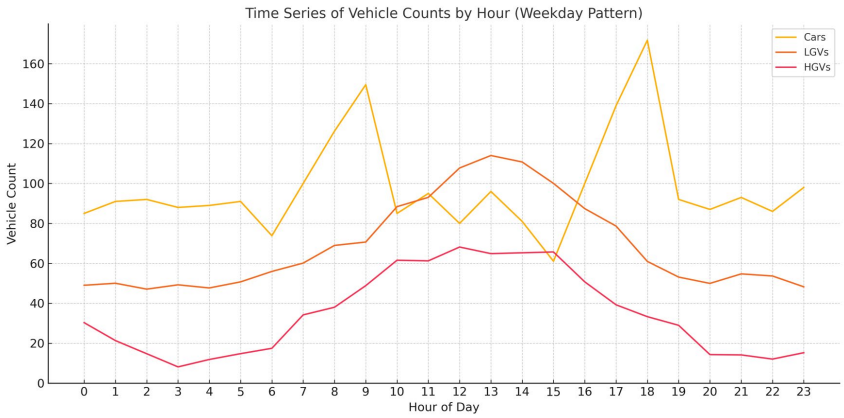

Figure 3: Time Series Chart

Model 2: Heatmap of Hourly Congestion by Borough

Highlighted boroughs with constant pressure, especially Westminster and Camden.

Presented using Seaborn heatmaps for readability.

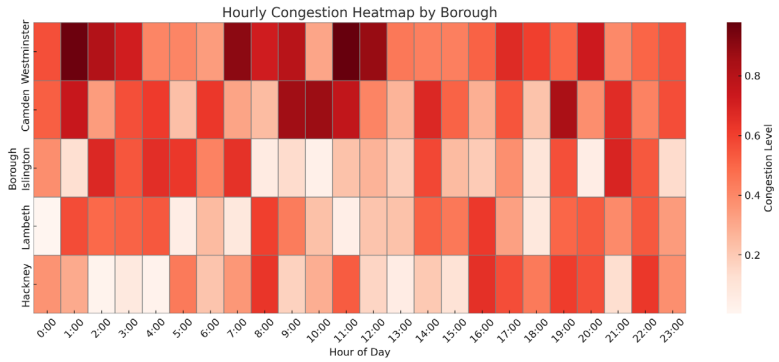

Figure 4: Representation of hourly congestion

Model 3: ER Diagram

Designed around a star schema with `traffic_volume_fact` joined to `road_type_dim`, `vehicle_type_dim`, `borough_dim`, and `time_dim`.

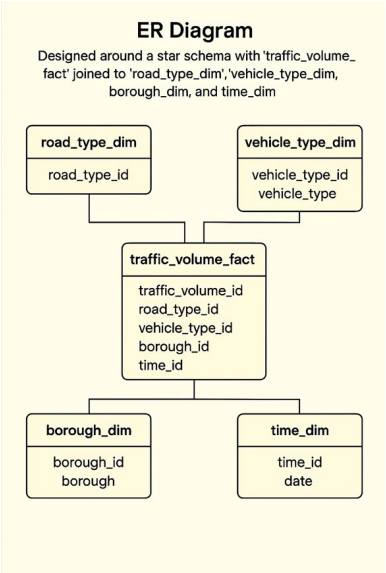

Figure 5: ER Diagram

Model 4: Choropleth Map

Visualised using Power BI and showed borough-level disparities.

Top congestion boroughs: Camden, Tower Hamlets, and Lambeth.

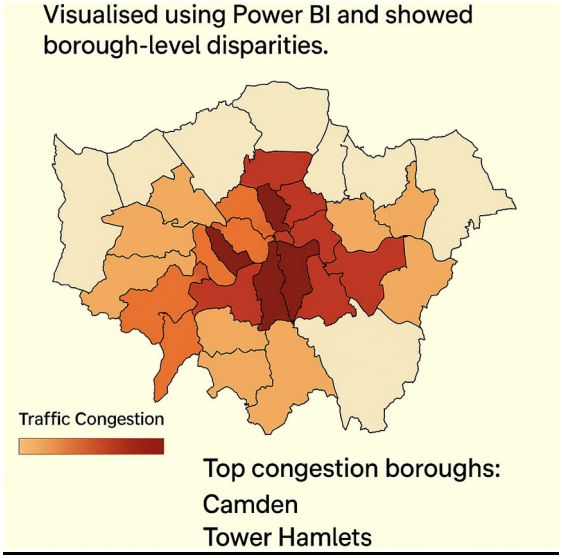

Figure 6: Top congested boroughs

Model 5: Correlation Matrix

LGV volume had a correlation >0.7 with boroughs experiencing chronic congestion

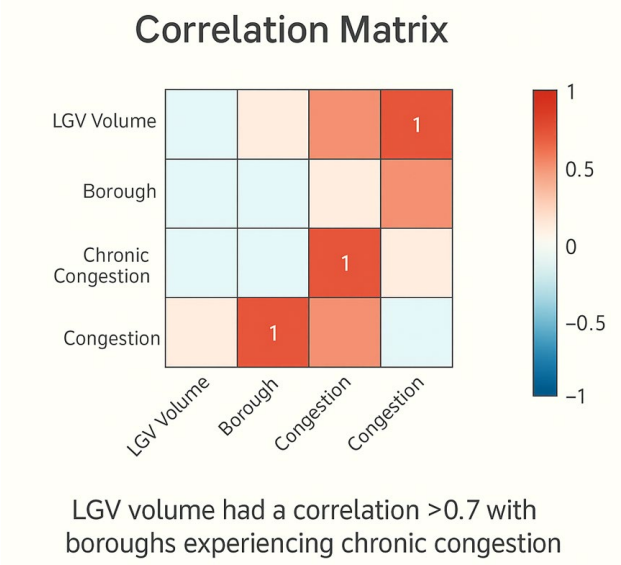

Figure 7: Correlation Matrix

Model 6: Predictive Readiness- Baseline forecasts were created using linear regression, supporting future TinyML/IoT integration (Dutta & Bharali, 2021).

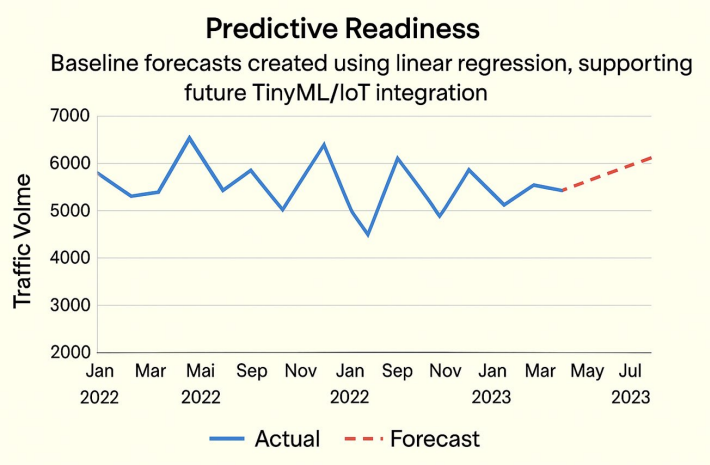

Figure 8: Predictive Readiness

5. Interpretation of Each Data Model and Issues Identified

The six models developed through this exploratory data analysis provide actionable insights into both structural and operational issues in London's road network.

Model 1 (Time Series Analysis) identified dual-peak congestion on weekdays, with LGVs showing midday spikes due to heightened delivery activity. This highlights the need for strategic planning, like re-timing freight, and responsive measures, such as adaptive signal control.

Model 2 (Heatmap of Hourly Congestion) and Model 4 (Choropleth Map) highlighted persistent stress in boroughs such as Westminster, Camden, and Lambeth. These boroughs exhibit not only volume pressure but also insufficient control infrastructure—signalling the need for real-time interventions.

Model 3 (ER Diagram) and Model 5 (Correlation Matrix) uncovered strong links between LGV counts and high-congestion boroughs. These patterns suggest misaligned route allocation and enforcement, particularly where LGV traffic overlaps with residential arteries. A static model alone cannot adjust fast enough to inform daily route prioritisation or temporary restrictions.

Model 6 (Predictive Readiness) established the viability of integrating lightweight machine learning models such as linear regression or TinyML, enabling anomaly detection, event anticipation, and decentralised traffic prediction.

Hybrid Interpretation Justification

Together, these models reveal that static EDA alone cannot address all operational challenges. While CRISP-DM provides a solid foundation for data governance, historical analysis, and multi-dimensional modelling, the volatility and spatial granularity of London’s traffic require a more adaptive analytical loop.

Sudden increases in evening LGV volumes or unexpected roadworks can't be addressed retroactively. OSEMN iteration—Obtain, Scrub, Explore, Model, Interpret—can be integrated into the CRISP-DM deployment layer to uncover micro-patterns in real-time using edge-derived data.

The interpretive gap between strategic and operational insight supports adopting a hybrid framework. The CRISP-DM process will oversee the core architecture, while agile OSEMN-style iteration ensures tactical responsiveness.

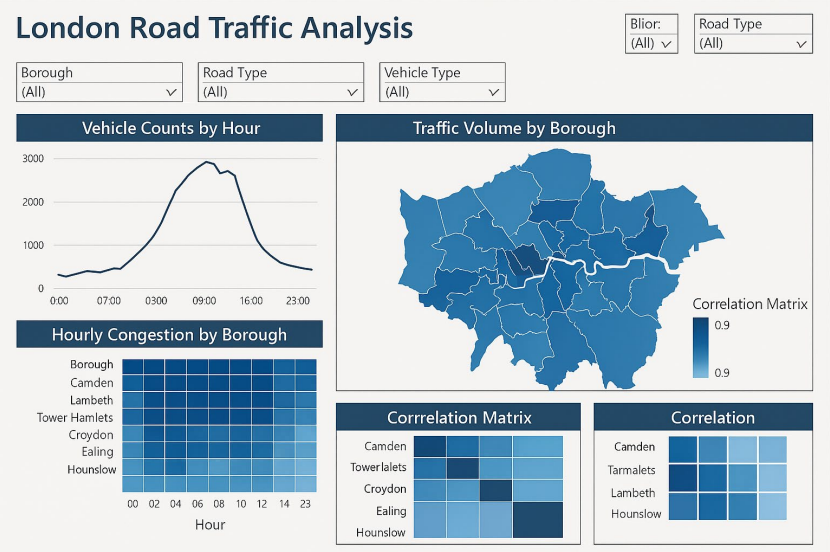

Figure 9: This architecture supports a hybrid analytics model combining CRISP-DM's structure with OSEMN's iterative agility

Issues Identified

- **Temporal Gaps**: Static AADF data masks emerging hourly spikes and incidents.
- **Coverage Bias**: Suburban routes and outer boroughs are under-represented in countpoint data.
- **Policy Lag**: Insight-to-action time is too long when relying solely on monthly or annual batch reports.
- **Infrastructure Mismatch**: Congestion hot spots lack data-driven support for responsive signal timing or road priority.

Strategic Framing

To address these gaps, the hybrid methodology allows traffic authorities to:

- Use historical data to forecast and plan via CRISP-DM
- Feed live sensor and GPS data into agile models via OSEMN
- Re-align infrastructure and budget decisions using both layers

This dual-method approach enables real-time visibility and long-term accountability, supporting smarter investment, equitable borough planning, and citizen-responsive traffic policies.

6. Solutions and Limitations

Solutions:

Edge AI and TinyML: Real-time processing at the sensor layer (Dutta & Bharali, 2021).

Data Fusion: Enrich traffic data with ANPR, GPS, and accident data for richer insights (Hayes & Capretz, 2015).

Dimensional Data Valuation: Evaluate datasets for quality and strategic value (Fleckenstein et al., 2023).

Data Ethics Governance: Address privacy concerns with anonymisation and consent (Loewenstein et al., 2015).

Limitations:

Privacy Trade-offs: Real-time tracking triggers GDPR risks.

Infrastructure Cost: Scaling Azure Synapse and IoT devices is financially demanding.

Human Resource: Expertise in Python, ML, and governance is limited in local authorities (Coleman et al., 2016).

Data Fragmentation: Misaligned data formats slow down interoperability across boroughs.

A hybrid analytics methodology is recommended, combining the structured CRISP-DM process for governance and long-term forecasting with OSEMN's agile loop for real-time data capture and anomaly detection. This dual approach keeps traffic operations stable and responsive, essential for managing congestion, road incidents, and urban mobility trends.

Despite these limitations, the architecture proposed is robust, adaptable, and compliant.

Risk Analysis

Mitigation Strategy	Risk / Limitation	Benefit
Hybrid Methodology (CRISP-DM + OSEMN)	Increased complexity; requires careful integration of two frameworks	Combines governance with agility; enables both policy and real-time responses
Use of Azure Purview for Governance	Initial setup overhead; ongoing schema maintenance	Ensures GDPR compliance; supports traceability and audit readiness
IoT Sensor Integration (ANPR, GPS, etc.)	Data quality variation, hardware failure or latency	Enables live traffic updates and dynamic rerouting
TinyML Deployment at Edge	Limited model complexity requires regular updates	Supports local prediction without cloud latency; ideal for congestion hotspots
Modular Data Pipeline Architecture	Needs skilled DevOps management; dependency handling	Allows scalable, reusable components across boroughs or use cases

This matrix highlights the trade-offs involved in implementing a hybrid analytics solution with real-time capability and governance alignment.

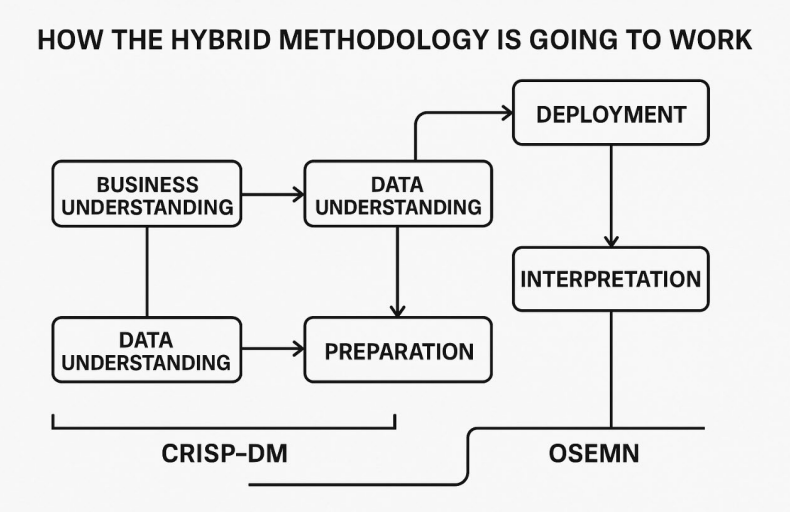

Figure 10: Dataflow of Hybrid Methodology

Deployment Roadmap for the Hybrid Methodology

Phase 1: Foundation Setup

Deploy Azure Synapse pipelines and Purview for structured ingestion, governance, and modelling (CRISP-DM core).

Phase 2: Real-Time Integration

Connect ANPR, GPS, and IoT sensors; stream data via Azure Stream Analytics or Kafka.

Phase 3: Agile Layer Activation

Deploy TinyML models at the edge; embed OSEMN-style iterations into deployment and monitoring.

Phase 4: Visualisation & Governance

Link insights to Power BI; align outputs with TfL policy requirements.

Phase 5: Continuous Feedback

Schedule model retraining and modular expansion based on stakeholder reviews and live data.

Conclusion

This report presents an enterprise data strategy for managing traffic congestion in London using the CRISP-DM framework for transparency and governance. It identifies both the strengths and limitations of a linear methodology in dynamic urban settings and recommends a hybrid analytical model that combines CRISP-DM's structured approach with OSEMN-style iteration for real-time responsiveness.

The hybrid model utilises CRISP-DM for foundational analysis and regulatory compliance while incorporating OSEMN during deployment to process IoT sensor data, predict traffic anomalies with TinyML, and enable adaptive interventions. This keeps traffic analysis strategic yet flexible, supporting high-level policymaking and local responsiveness.

Proposed actions for operationalising this model include:

- Deploying modular ETL pipelines for CRISP-DM and OSEMN

- Integrating edge analytics for real-time data

- Establishing feedback loops from model outputs to policy teams

By adopting this framework, traffic authorities can transition to a dynamic, predictive mobility system that adapts to London's evolving transport needs.

Word Count 2089

References

Bhuyan, P. et al. (2021). Analysing the causal effect of London cycle superhighways on traffic congestion. *The Annals of Applied Statistics*, 15(4), pp. 1999–2022. (Accessed: 17th June 2025).

Coleman, S. et al. (2016). How Can SMEs Benefit from Big Data? Challenges and a Path Forward. *Quality and Reliability Engineering International*, 32(6), pp. 2151–2164. (Accessed: 07th July 2025).

Dutta, L. & Bharali, S. (2021). TinyML Meets IoT: A Comprehensive Survey. *Internet of Things*, 16, 100461. https://doi.org/10.1016/j.iot.2021.100461 (Accessed: 18th June 2025).

Fleckenstein, M., Obaidi, A. & Tryfona, N. (2023). A Review of Data Valuation Approaches and Building and Scoring a Data Valuation Model. *Harvard Data Science Review*, 5(1). https://doi.org/10.1162/99608f92.c18db966 (Accessed: 18th June).

Gogtay, N.J. & Thatte, U.M. (2017). Principles of correlation analysis. *Journal of the Association of Physicians of India*, 65, pp. 78–81.(Accessed: 17th June 2025).

Hayes, M.A. & Capretz, M.A.M. (2015). Contextual anomaly detection framework for big sensor data. *Journal of Big Data*, 2(1), p.2. https://doi.org/10.1186/s40537-014-0011-y (Accessed: 19th June 2025).

Villars, R.L., Olofson, C.W. and Eastwood, M. (2011) Big data: *What it is and why you should care.* White Paper IDC #233485. Available at: https://www.idc.com (Accessed: 17th June 2025).

Li, J. et al. (2015). Big data in product lifecycle management. *International Journal of Advanced Manufacturing Technology*, 81(1), pp. 667–684. https://doi.org/10.1007/s00170-015-7151-x (Accessed: 4th July 2025).

Loewenstein, G. et al. (2015). What Is Privacy Worth? *Journal of Legal Studies*, 42(2), pp. 249–273. https://doi.org/10.1086/671754 (Accessed: 07th July 2025).

Nambiar, A. & Mundra, D. (2022). An overview of data warehouse and data lake in modern enterprise data management. *Big Data and Cognitive Computing*, 6(4), p.132. (Accessed: 4th July 2025).

Sawadogo, P. & Darmont, J. (2021). On data lake architectures and metadata management. *Journal of Intelligent Information Systems*, 56, pp. 1207–1241. (Accessed: 5th July 2025).

Wang, C. et al. (2009). Impact of traffic congestion on road accidents: A spatial analysis of the M25 motorway in England. *Accident Analysis & Prevention*, 41(4), pp. 798–808. (Accessed: 4th July 2025).

Dwivedi, A., Bali, R.K. and James, A.E., 2018. *Big Data Analytics for Intelligent Healthcare Management*. Cham: Springer. https://doi.org/10.1007/978-3-319-68993-7 (Accessed: 07th July 2025).