HPDM099: Midpoint Project Review

**Project:** Forecasting NHS Workforce Attrition and Retention Drivers

**Host Organisation:** Somerset NHS Foundation Trust

**Supervisors:** Michal Kozaczuk (Host), Neil Vaughan (Academic)

# Summary of Progress to Date

The primary phase of this project focused on data access, familiarisation with the datasets and establishing a foundation for a scalable analytical workflow. Once datasets such as NHS\_Leavers\_Detail, NHS\_Starters\_Detail, Staff\_List and other documents became accessible, I was able to start data wrangling, guaranteeing that missing data are handled properly, standardising field names and reconciling formats across tables.

An early achievement was building a reproducible and clean data pipeline in Python. This involved temporal aggregation of monthly leavers and starters, imputation strategies for missing demographic attributes and consistency checks such as ensuring persistent employee numbers across multiple assignments. I extended this to exploratory analysis, creating time series decomposition and visualisations of turnover by staff group, service line and demographic segments using matplotlib, seaborn and pandas.

cleaning steps included:

* Standardising date format – Effective Date, Leaving Date
* Binning Length of Service (LOS) to fixed intervals
* Calculating net movement using shared keys across Staff List and Leavers
* Removing duplicate employee entries based on persistent Employee Number and filtering internal assignments

Time series forecasting by setting a baseline using auto\_arima from pmdarima library. ARIMA(0,0,1)(0,0,1,12) AND SARIMAX(3x1x0)x(1,0,0,12) were fitted, highlighting both trend and seasonal components. Rolling forecast cross validation and using a 6 step horizon was conducted to assess model robustness, yielding a mean absolute error (MAE) of 2.73. Diagnostics on residuals showed approximate whiteness with minor autocorrelation remaining at seasonal lags.

The modelling pipeline intergrates:

* Auto\_ARIMA seasonal model with RFCV
* Forecast visualisation with prediction intervals
* Aggregated monthly metrics adjusting for varying month lengths

Above all, I have created the groundwork for two novel components:

1. **Internal Movers Detection:** By leveraging the Staff\_List dataset’s stable employee number (but changing assignment IDs), I engineered the code to detect internal movements that bypass ‘leavers’ and ‘starters’ datasets. I visualised these via interactive network graphs (using NetworkX and Pyvis) and bubble plots (plotly express), highlighting intra-organisational flow hotspots. These tools now reveal latent movement patterns previously invisible to senior HR stakeholders.
2. **Hierarchal Forecast Planning:** In consultation with one of the data scientists from the host organisation Chris and referencing ‘Forecasting: Principles and Practices (Pythonic Way), I identified grouped time series forecasting as an effective next step. I am now transitioning toward Nixtlaverse’s StatsForecast and HierarchalForecast libraries, defining nested hierarchies (Staff Group > Service Group> Organisation > Department) to forecast attrition at different organisational levels.

# Projected Outputs Based on The Currently Position

Based on the current course and progress:

* **Validated ARIMA forecasts** for total NHS leavers (with diagnostics and confidence intervals)
* **Interactive Dash Plotly dashboard**, including:
  + Monthly leavers/starters
  + Net change calculations
  + Forecast toggles
  + Filter controls for demographic attributes
* **Hierarchical forecast models** using Nixtla to support grouped projections ( e.g, by Staff Group/Service).
* **Internal movers network graphs** visualised via Pyvis, featuring:
  + Node based timelines from HIRE 🡪 internal reassignments 🡪 LEAVE
  + Edge tooltips showing department/service/time
  + Interactive filtering per month
* **Final report and technical appendix** documenting pipeline, models and dashboards.
* **GitHub repository** of reproducible code and a README for deployment.

# Changes to Original Project Plan

While the primary goals remain the same, a few modifications improved project direction, these adaptations strengthen and align with personal and stakeholder needs while maintining deliverables and timeline.

Modifications:

1. **Forecasting Methodology Shift:** Originally planned to use Prophet only. After diagnostic limitations, ARIMA was used for benchmarking and now moving to Nixtlaverse for scalability and hierarchical support.
2. **Internal Movers Module:** Not originally planned but developed due to insight gap in leaver/starter records. It became a distinct analytical asset.
3. **Dashboard Re-Architecture**: Initial dashboard attempts prioritised frontend aesthetics. Upon feedback, I restructured to ensue backend logic is stable first using modularised callbacks and reusable Dash components.
4. **Increased Data Span**: To address insufficient seasonal cycles (only up to 13 months initially), additional historical data was utilised to extend forecasting reliability.

## Updated GANNT Chart

A graph with different colored bars

Description automatically generated with medium confidence

# Reflections on Experience and Learning

This project has been extremely enjoyable and academically challenging. Navigating real world NHS workforce data exposed me to the intricacies of data quality, temporal continuity and metadata ambiguity, issues I rarely face in clean academic datasets. It challenged my technical fluency in wrangling longitudinal assignments, inferring patterns (like internal mobility) and matching analytical decisions with operational meaning.

Building statistical models and validating them with diagnostics strengthened my confidence in interpretable forecasting, particularly balancing a model fit with stakeholder usability. The introduction to Nixtlaverse has been a turning point, showing how advanced packages can support real time, grouped forecasts at scale.

The experience also improved my communication skills, translating technical tasks (such as network analysis logic) to Workforce professionals during meetings whole while coordinating outputs with actionable workforce insights.

All in all, moments of technical setback (Dash bugs, data gaps etc.) helped me build resilience, debugging discipline and patience. This reflective learning process ensures not only a successful MSc result but a strong basis for future data science practice in healthcare settings.

# Items for Discussion at Review

* Finalising the hierarchical structure for Nixtlaverse forecast
* UX feedback for Dash dashboard: prioritising filters vs toggles
* How best to link network graph into dashboard
* Deployment support options
* Review on writing format: Should the logic be explained in full in final report appendix, should the internal movers graph be included?

# Summary Conclusion

To date, the project is progressing well, with a strong technical foundation, comprehensive documentation and stakeholder engagement. The ARIMA and time series models act as validated baselines and work is now transitioning to higher level forecasting with group hierarchies using Nixtla libraries. Alongside this, the network visualisation prototype for internal movers displays this project as uniquely tailored to the NHS context.

By continuing to focus on robust modular code, clear group structures and interactive visuals, the project is well positioned to deliver actionable insight for Somerset NHS Foundation Trust and contribute a replicable framework for NHS workforce.

# Appendix A: A copy of original proposal

A document with text and a list

Description automatically generated with medium confidence

# Appendix B: Original Project plan

Initial Project Plan – MSc Health Data Science (HPDM099)

Project Title: Forecasting NHS Workforce Attrition and Retention Drivers

Supervisor: Michal Kozaczuk

Academic supervisor: Neil Vaughan

Background & Aim

The NHS seeks to understand and predict staff turnover (“leavers”) and net headcount changeto inform retention strategies. By leveraging the Senior HR Dashboard data (April 2025) andtime series forecasting models, including Prophet and advanced Nixtlaverse libraries, this

project will:

- Quantify historical monthly leavers and starters

- Forecast attrition for FY 2025-26

- Identify high – risk demographic and service/staff groups

- Deliever an interactive dashboard for stakeholders

Objectives

1. Data ingestion & cleaning: prepare leavers and starters data with demographic filters (service group, staff group, gender, ethnicity, disability, nationality, age band,recruitment source, length of service)
2. Time series modelling: develop and validate forecast for leavers, starters and net change using Prophet; benchmark with StatsForecast for scalability.
3. Dashboard development: build a Dash/Plash app enabling interactive filtering and scenario analysis.
4. Evaluation: perform cross validation, compute MAE/RMSE and confidence interval coverage, refine models.
5. Dissemination: produce final report, code repository and user guide.

Methodology & Deliverables

- Review → survey of workforce forecasting, retention factors

- Data pipeline → CSV ingestion, cleaning LOS binning → clean pandas dataframe

- Exploratory analysis → trends by group, seasonality → EDA notebook

- Forecast models (Prophet functions, Nixtlaverse comparison) → Model scripts + performance

- Validation → cross validation, metrics report → Deliverable: CV results + recommendations

- Dashboard → DASH app with 9 filters, time series plots -→ running app code and demo link

- Report and presentation → draft chapters and slidesA white sheet with black text

Description automatically generated

Resources

Software: Python, JupyterLab, Git, Dash-Plotly

Libraries: Pandas, Prophet, Nixtlaverse (StatsForecast, NueralForecast), scikit-learn,

matplotlib

Data: NHS Senior HR Dashboard (leavers + starters 2025)

Local workstation/environment

Risks/Uncertainties

- Data quality: Missing or inconsistent demographic fields → mitigate with imputation/

potential exclusion

- Model performance: Seasonal anomalies or external shocks – fix with holiday

regressors/ external covariates

- Technical integration: Dash deployment complexities → plan early prototypes

- Time constraints: balancing coding, analysis and write up – keep up with weekly/biweekly reviews

Next steps

- Share plan with supervisors and review

- Delve further into planning and prototyping

# Appendix C: Copies of weekly log reports

A screenshot of a black screen

Description automatically generated

A black and white document with white text

Description automatically generated

A black and white document with text

Description automatically generated

A black and white document with text

Description automatically generated

A screenshot of a computer screen

Description automatically generated

A screenshot of a black screen

Description automatically generatedA screenshot of a black and white page

Description automatically generated