1. Background & Context

The National Health Service (NHS) in England faces pressure balancing clinical capacity with rising patient demand amid demographic shifts and post pandemic healthcare expectations. General practitioners must manage appointment availability, ensure high attendance rates, and optimise resource utilisation, all while remaining financially sustainable under increasingly constrained budgets. Missed appointments (DNA) impose avoidable costs estimated at £216 million annually and disproportionately affect vulnerable patients who face barriers to healthcare access. Simultaneously, external signals such as public sentiment on social media platforms like Twitter offer real time insights into patient experiences and system strain, enabling proactive capacity management. This report examines 30 months of wrangled NHS data spanning appointments across regions, durations, and service categories, alongside a representative sample of Twitter data, to answer whether current capacity and staffing levels suffice and how resources are actually utilised in practice.

2. Analytical Approach

Data Ingestion & Initial Exploration

We imported three cleaned data sources using Pandas with error handling:

appointments_regional.csv → ar (regional appointment volumes)

actual_duration.csv → ad (consultation timing data)

national_categories.xlsx → nc (service classification data)

A sample tweets.csv was also loaded to explore external sentiment patterns. We applied pd.to_datetime with error='coerce' to parse date strings (appointment_month, appointment_date), ensuring consistent temporal alignment, and pd.to_numeric with downcast='integer' to optimise memory usage for duration fields. Comprehensive metadata inspection using .info(), .describe(), and .value_counts() revealed systematic missing values, anomalous durations (< 1 min or > 60 min flagged as "Unknown" likely due to system recording errors), and inconsistent "Unknown" categories in appointment_mode and appointment_status requiring careful imputation strategies.

Data Wrangling & Aggregation

We performed rigorous, consistent wrangling steps for each DataFrame to ensure analytical validity:

Temporal filtering to align date ranges across sources, restricting multi table joins to December 2021 – June 2022, representing the period when all data sources demonstrate complete coverage and reliability.

Strategic grouping by key dimensions via groupby operations with .sum() aggregation, followed by .reset_index() to create analysis ready tidy tables:

ar_monthly: total appointment volumes per month

ar_agg: granular monthly breakdown by healthcare professional (HCP) type, appointment status, delivery mode, and booking lead-time interval

nc_ss: monthly service totals categorised by service_setting classification

ad_dur: monthly consultation duration statistics including average, median, and distribution percentiles.

Computing derived performance measures:

Daily appointment average = monthly total ÷ 30 days

System capacity utilisation = (daily average ÷ 1,200,000 daily capacity) × 100%, based on NHS England's efficiency plan.

Advanced text processing on tweets dataset: generating frequency distributions using Pandas Series.value_counts(). We examined tweet_retweet_count and tweet_favorite_count distributions for potential engagement based weighting schemes.

Visualisation & Reproducibility Framework

All visualisations leveraged Seaborn with Matplotlib backend for graphics. For time series analysis, we formatted monthly periods as YYYY-MM strings optimising axis readability. We applied line plots for temporal trend analysis, boxplots for cross sectional distribution comparisons, and horizontal bar plots for hashtag frequency rankings. Configuration included standardised figure dimensions (15×12 inches), and whitegrid style for maximum clarity. All code cells include comprehensive Markdown documentation adhering to PEP 8 naming conventions. Version controlled Jupyter Notebook backup ensures full analytical reproducibility.

Critical Limitations & Quality Considerations

Uneven coverage: Practice participation rates vary systematically, monthly coverage fluctuates between 85-98%, necessitating weighted adjustments using available coverage correction files.

Systematic data quality issues: TPP SystmOne electronic health records demonstrated significant DNA under reporting during June to November 2018 transitional period; Cegedim Vision system users consistently lacked reliable appointment_mode classification.

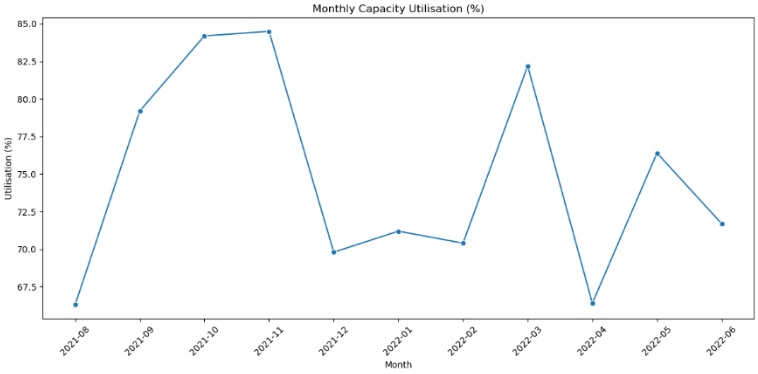
Social media sampling constraints: Twitter dataset represents geographically and demographically limited sample without comprehensive sentiment analysis or verified user authenticity.

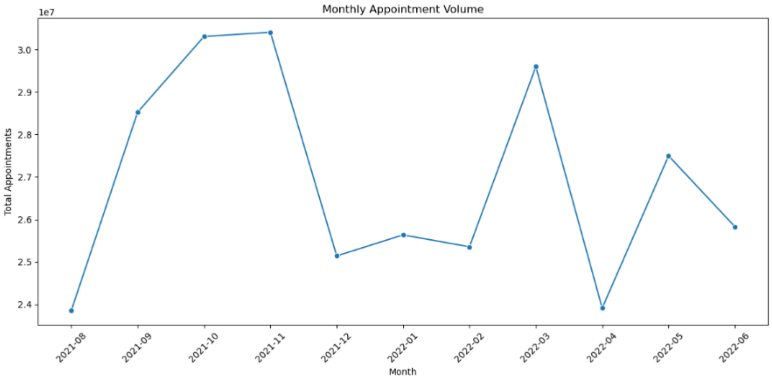
Duration measurement inconsistencies: Different clinical system suppliers record consultation start and end timestamps using varying methodologies, introducing systematic measurement bias across practice networks.

3. Visualisation & Insights

Capacity & Demand Alignment Patterns

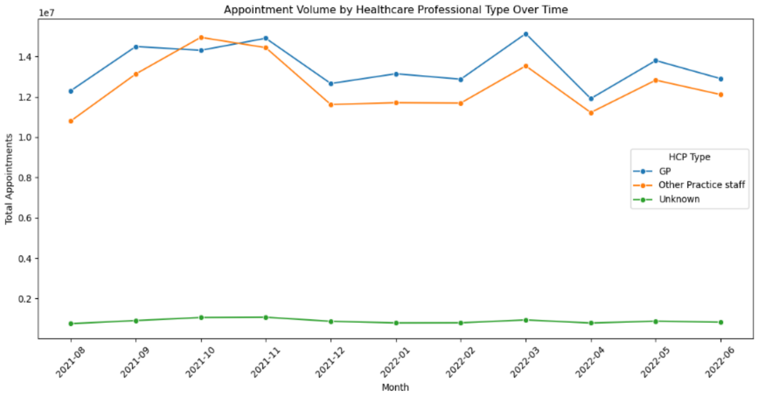
A dual axis line plot revealing monthly appointment volumes against calculated capacity utilisation exposed predictable seasonal demand cycles: volumes increased progressively from 24 million appointments in August 2021 to peak demand of 30 million during November 2021, followed by a surge reaching 29 million in March 2022. System utilisation fluctuated between 70-90% during typical periods, approaching 85% at seasonal peaks. Complementary boxplots confirmed narrow interquartile ranges in monthly appointment totals except during predictable holiday periods.

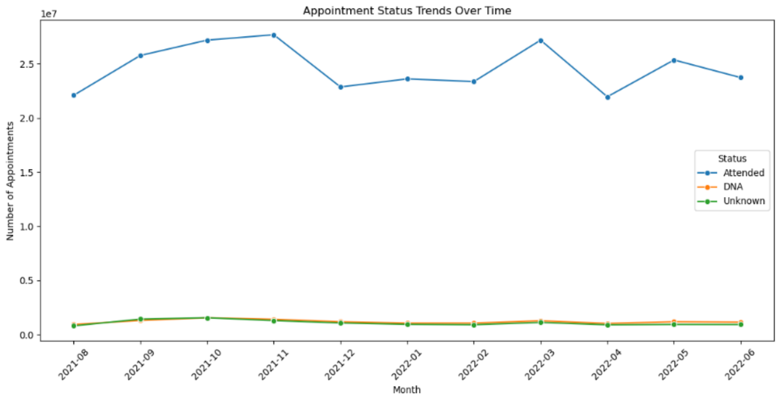




Professional Workforce Mix & Attendance Dynamics

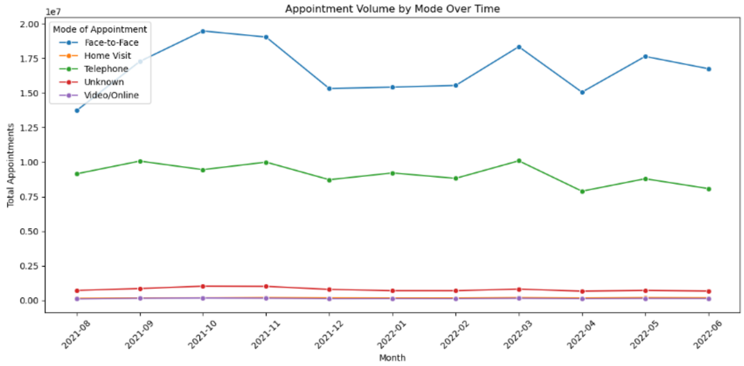
Line plots ranked by hcp_type demonstrated General Practitioners (GP) consistently handling approximately 55% of total appointment volume, with Other Practice Staff (nurses, healthcare assistants, pharmacists) scaling proportionally during demand surges. Attendance pattern analysis indicated remarkably stable 94-97% attendance rates with persistent DNA baseline around 3% and Unknown status classifications at 2%, notably remaining constant despite significant demand fluctuations.

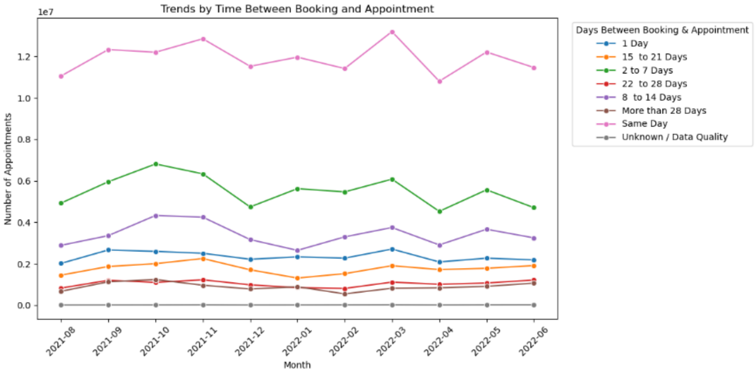




Appointment Delivery Modes & Booking Lead Times

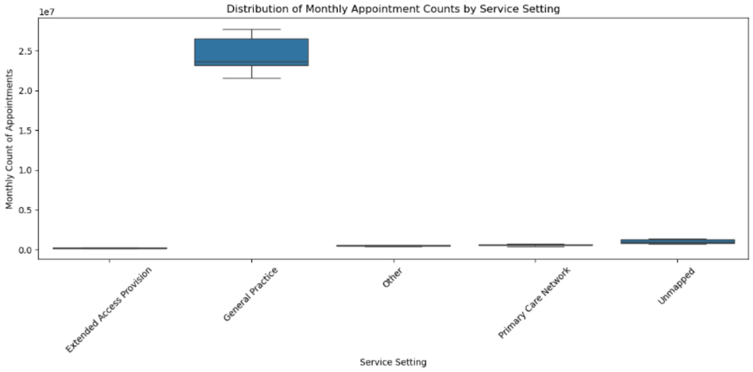
Comprehensive analysis of appointment_mode preferences confirmed Face-to-Face consultations dominating (16-20 million monthly) alongside substantial Telephone adoption (8-10 million monthly); Online/Video consultations and Home Visits remained consistently marginal under 200,000 monthly. Booking to appointment interval distributions clustered predictably at same day access (12 million monthly) and 2-7 day advance booking (6 million monthly), maintaining stability across observation period.

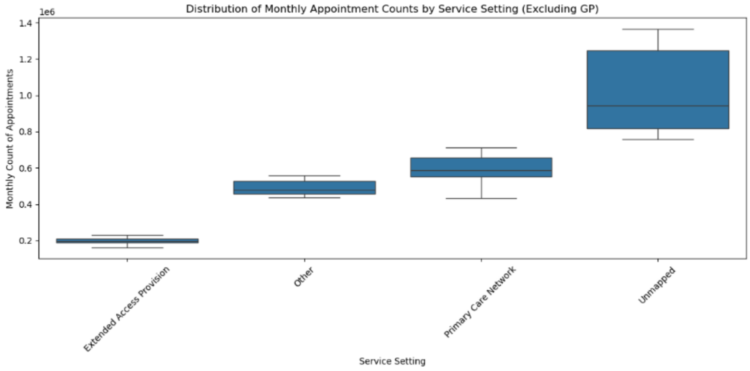




Service Setting Classification Analysis

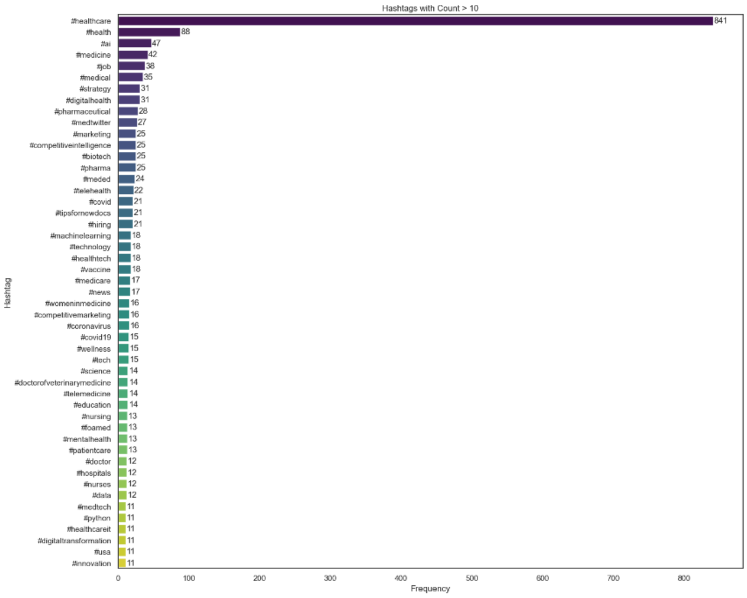
Detailed boxplots comparing service_setting categories illustrated General Practice's overwhelming volume dominance (24 million monthly) versus emerging Primary Care Network services (0.6 million monthly), Extended Access provisions (0.2 million monthly), and miscellaneous "Other" categories. The substantial Unmapped classification exhibited concerning variability (0.8-1.4 million monthly), highlighting significant data quality gaps requiring urgent attention.

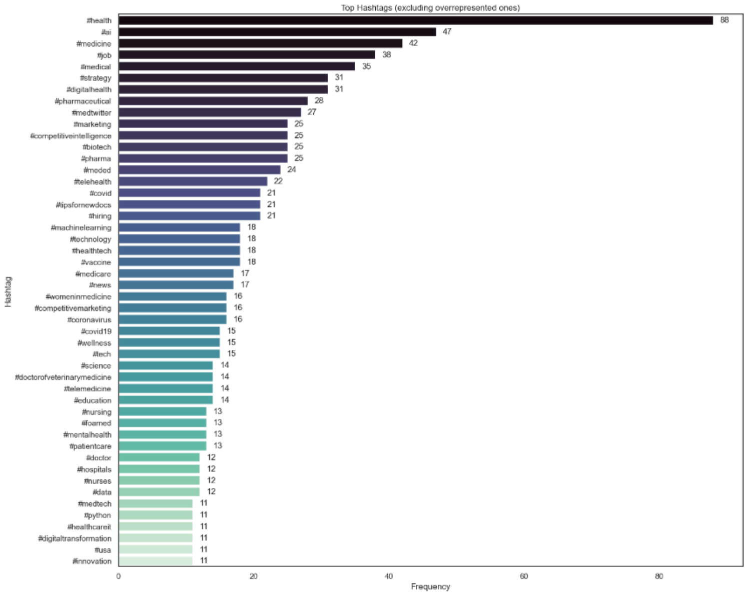




External Social Media Signal Detection

Hashtag frequency analysis (filtered for count>10, excluding generic terms) identified predominant NHS related discourse themes (#healthcare, #ai, #medicine, #job), with engagement metric distributions showing characteristic heavy right skewing patterns. This suggests implementing engagement weighted hashtag prioritisation could effectively surface emerging public health concerns and system pressures.





4. Patterns & Predictions

Seasonal Demand Forecasting: Autumn (October-November) and pre-Easter (March) periods demonstrate predictable high demand patterns. Summer and Easter intervals provide essential system recovery capacity.

Persistent Attendance Challenges: DNA rates represent consistent 3% baseline requiring targeted behavioural interventions rather than reactive seasonal responses.

Digital Health Transformation Opportunities: While telephone consultations scale naturally with demand, video consultations remain significantly underutilised. Strategic tele-health expansion could unlock substantial capacity relief.

Adaptive Workforce Management: Other Practice Staff demonstrate excellent demand responsiveness mirroring GP patterns, suggesting enhanced clinical role expansion during peak periods could significantly alleviate GP workload pressures.

Data Infrastructure Investment Priorities: Resolving "Unmapped" service setting classifications and duration recording inconsistencies will substantially improve forecasting model accuracy.

Integrated Predictive Analytics: Combining real time social media sentiment analysis with historical booking patterns and capacity utilisation metrics enables proactive staffing allocation and resource optimisation.

Future Research Recommendations: Deploy sophisticated time series forecasting models (ARIMA) that implement automated sentiment analysis pipelines; conduct comprehensive geospatial analysis identifying regional capacity disparities.