Phil Best

# NHS Capacity Analysis: Technical Report

## 1) Problem Statement

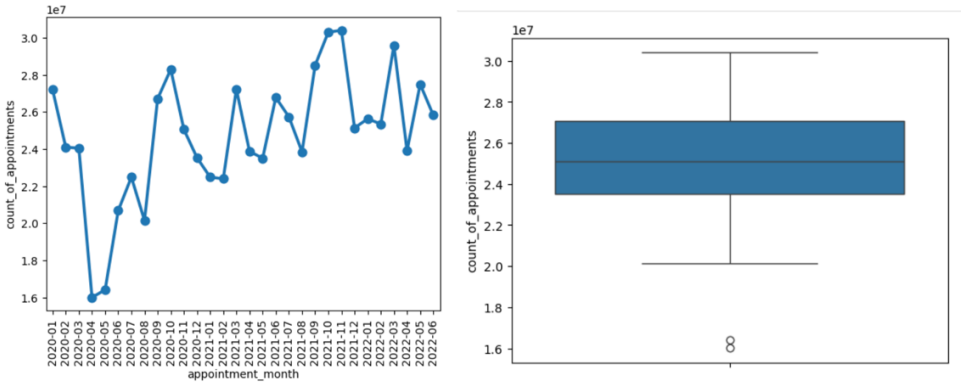
Analysis into whether the NHS has enough capacity to meet demand, and how resources are currently being utilized. Failing to understand this could lead to poor investment decisions and ultimately, poorer provision of healthcare amongst the UK population.

## 2) Collecting the data [6] to [9]

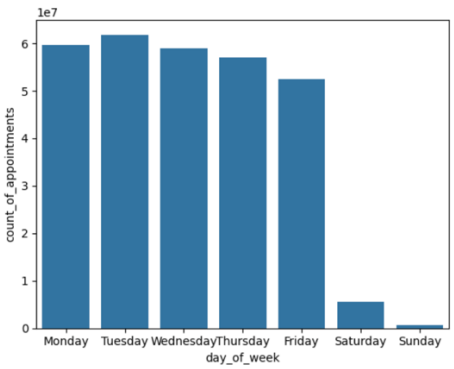
- Imported 'actual_duration.csv' and 'appointments_regional.csv' using pd.read_csv(), assigning the names 'ad' and 'ar' to the resulting data frames
- Imported the ''national_categories.xlsx' using pd.read_excel(), assigning the name 'nc' to the resulting data frames

## 3) Clean the data [12] to [31]

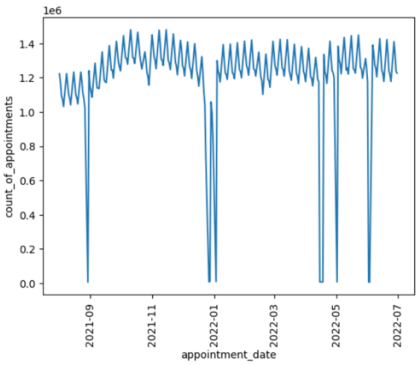
- Checked for missing values across ad, ar and nc using isna().
  - No missing values found
- Checked data types using df.dtypes to show inconsistencies or inaccurate data descriptions
  - ad - Converted appointment_date to datetime
  - ar - Converted appointment_month to date from string
  - nc - Converted appointment month to date from string and appointment_date to datetime
- Further cleaned data during '5. Analyse the data' phase
  - Outlier analysis completed with lineplot and boxplot identifying Covid-19 impact, leading to removal of data from analysis before August 2021 **[74] + [75]**

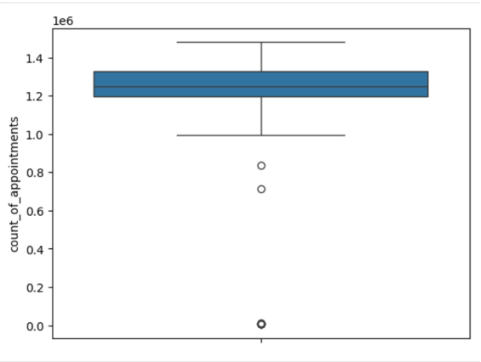



  - Outlier analysis completed on appointment_count per day of week, identifying weekends as outliers **[86] + [87]**



- Outlier analysis completed on weekdays, removing any days under 1,000,000 as per the interquartile range **[93] + [94]**




- Removed spaces from time_between_booking_and_appointment through grouping and renaming the rows via a dictionary **[156]**

**4) Explore the data [34] to [68]**

- Explored each dataframe using df.info() and meta data file to identify the Dimensions and Measures (Appendix 1)
  - Location and Date data consistent across datasets (can be used as Primary keys)
  - Only measure is count_of_appointments

- Determined the counts and hierarchy of location data using .nunique() **[34] to [36]**
  - Regions: 7
  - ICB Locations: 42
  - Sub ICB Locations: 106

- Checked start and end dates of data in each data frame using .min() and .max() **[38] to [40]**
  - ad: 1st Dec 2021 to 30th June 2022 (Daily data)
  - nc: 1st Aug 2021 to 30th June 2022 (Daily data)
  - ar: January 2020 to June 2022 (Monthly data)

- Checked consistency of appointment counts by month across each data frame by grouping the datasets by appointment month using .group() and summing appointment_count **[41] to [51]**
  - ar and nc datasets have consistent appointment counts per month and therefore can be merged if required
  - ad has inconsistent counts with other datasets

ar data

| | appointment_month | count_of_appointments |
|---|---|---|
| 0 | 2021-08 | 23852171 |
| 1 | 2021-09 | 28522501 |
| 2 | 2021-10 | 30303834 |
| 3 | 2021-11 | 30405070 |
| 4 | 2021-12 | 25140776 |
| 5 | 2022-01 | 25635474 |
| 6 | 2022-02 | 25355260 |
| 7 | 2022-03 | 29595038 |
| 8 | 2022-04 | 23913060 |
| 9 | 2022-05 | 27495508 |
| 10 | 2022-06 | 25828078 |

nc data

| | month_name | count_of_appointments |
|---|---|---|
| 1 | August | 23852171 |
| 10 | September | 28522501 |
| 9 | October | 30303834 |
| 8 | November | 30405070 |
| 2 | December | 25140776 |
| 4 | January | 25635474 |
| 3 | February | 25355260 |
| 6 | March | 29595038 |
| 0 | April | 23913060 |
| 7 | May | 27495508 |
| 5 | June | 25828078 |

ad data

| | month_name | count_of_appointments |
|---|---|---|
| 1 | December | 22853483 |
| 3 | January | 23597196 |
| 2 | February | 23351939 |
| 5 | March | 27170002 |
| 0 | April | 21948814 |
| 6 | May | 25343941 |
| 4 | June | 23715317 |

- Checking the number and names of categories for remaining dimensions through using .nunique() and .value_counts() **[53] to [68]**

**5) Analyse the data [72] to [236]**

- Posed analytical questions for each of the dimensions to identify any insights or trends
  a) How do the number of appointments change per month? **[72] to [78]**
  b) How do the days of the week influence appointment count? **[81] to [103]**
  c) How does weekday appointment count change through the year? **[106] to [112]**
  d) Which health professionals take the appointments? **[115] to [123]**
  e) What percentage of appointments are attended? **[126] to [133]**
  f) What percentage of appointments are made face-to-face? **[136] to [143]**
  g) Does appointment waiting time change over time? **[146] to [162]**
  h) Which service settings have the most appointments? **[165] to [174]**
  i) How many appointments were non-care related? **[177] to [186]**
  j) Which National Categories do most appointments fall into? **[189] to [200]**
  k) How do the number of daily appointments change between regions? **[203] to [211]**
  l) What insights can be gained from the Twitter data? **[214] to [236]**

- Followed the same approach for most of the above questions:
  - Filtered outliers as per above in '3. Clean the data'
  - Grouped data by date using .groupby() and summed appointment count
  - Where monthly data, added column showing number of days in each month to work out daily averages
  - Created exploratory visualisations using sns.lineplot or barplot
  - Pivoted the data to allow for total appointment count by month (using .pivot_table) and then added totals column with .sum()
  - Converted values in pivot table to percentages by dividing data by totals
  - Plotted percentages on a stacked barchart using .plot(kind='bar', stacked=True)

**6) Visualisations and Insights [239] to [282]**

- From the above exploratory analysis, identified key patterns that would answer:
  - Has there been adequate staff and capacity in the network?

  a) Covid significantly impacted available data **[239] to [243]**
  b) Overall, average daily usage is under the 1,200,000 capacity for period 08-2021 to 06-2022 **[245] to [250]**
  c) However there are large variances on different days of week, with all weekdays apart from Friday being over capacity **[252] to [258]**
  d) The average weekday is over capacity in every month other than August **[260]**

  - How resources are currently being utilized

  e) 90% of appointments occur in General Practice **[262]**
  f) 52% of appointments are handled by GPs **[264]**
  g) This means GPs are doing over the recommended max 25 appointments per day **[267] to [272]**
  h) There is a significant spike in 'Other Practice staff' appointments in October **[274]**
  i) This appears to be driven by uplift in Planned Clinical Procedures at this time (likely due to Flu Vaccine) **[276] to [277]**
  j) 4%-5% are appointments are lost through no-shows **[279] to [280]**
  k) Twitter data provided limited insight as it was unstructured and inconsistent **[282]**

- Visualisations chosen in-line with best practice (FT Visual Vocabulary). Change over time shown on linegraphs, categorical data on barcharts and percentages on stacked barcharts

- Colours and fonts in-line with NHS brand guidelines
- Enhanced analysis through incorporating other data sources:
  - BMA data on maximum number of appointments a GP can do per day
  - BMA analysis on number of GPs currently in workforce
  - NHS Flu vaccination roll-out dates

**7) Recommendation Summary**

- Business Recommendations
  - Capacity needs to be increased by an estimated **2,640 FTEs** to meet average weekday demand.
  - Incentivize GPs to open over weekends to reduce pent-up demand which is amplifying capacity issues
  - Start annual vaccine programs as early as possible to reduce appointments at busiest time of year
  - Explore measures to reduce unattended appointments such as financial penalties or incentives for attending
  - Before investment in further Twitter data, use NHS account for customer service, promoting own hashtags to aid analysis

- Analytical Recommendations
  - Create clear standardization of data entry to improve data quality
  - Further exploration on retention of GPs required (September 2015 -July 2022, a decrease of 6.3% of the GP workforce*)
  - Further exploration into regions and geographies, with direction on capacity for each area

## Appendix 1

| Dimensions | | | | |
|---|---|---|---|---|
| Category | Dimension Name | Data Source | Raw Data format | Data Transformation required? |
| Location | Icb_ons_code (ar) | ar, nc, ad | String | |
| | sub_icb_location_name | nc, ad | String | |
| | sub_icb_location_code | ad | String | |
| | sub_icb_location_ons_code | ad | String | |
| | region_ons_code | ad | String | |
| DateTime | appointment_date | nc, ad | nc: DateTime<br>ad: String | Change ad to DateTime format |
| | appointment_month | ar, nc | ar: String<br>nc: String | Change ad to DateTime format |
| Appointment context | context_type | nc | String | |
| | national_category | nc | String | |
| Attendance | appointment_status | ar | String | |
| Appointment method | appointment_mode | ar | String | |
| Appointment Type | hcp_type | ar | String | |
| | service_setting | nc | String | |
| Booking efficiency | time_between_book_and _appointment | ar | String | Rename columns |
| Appointment efficiency | actual_duration | ad | String | |

| Measures | | | | |
|---|---|---|---|---|
| Category | Measure Name | Data Source | Raw Data format | Data Transformation required? |
| # of appointments | count_of_appointments | ar, nc, ad | Integer | |