Diagnostic analysis using Python

Antonio Arantes

May 2024

Index

1. Background of the business scenario

The National Health Service (NHS) must make informed decisions about resource allocation and potential infrastructure expansion to match increasing population demands. Your data analyst team has been tasked with analysing internal NHS data to understand service utilization, missed appointments, and staff capacity. By investigating these areas, the NHS aims to determine if the current capacity is adequate or if adjustments are needed to improve efficiency and reduce costs associated with missed appointments. This analysis will provide actionable insights to guide budget allocation and optimize the use of existing resources.

2. Analytical approach

The team has been provided with three datasets: *actual_duration* (AD), *appointments_regional* (AR), and *national_categories* (NC). The first step after importing the datasets was to sense-check the data.

The three data frames did not have missing values. The next step was to investigate duplicates in the data. AD and NC did not have duplicate rows, while AR showed a different case with 21604 duplicated rows that needed further inspection.

[27]:
```
# Check one duplicate example.
de1 = ar[
    (ar['icb_ons_code'] == 'E54000008') &
    (ar['appointment_month'] == '2020-01') &
    (ar['appointment_status'] == 'Attended') &
    (ar['hcp_type'] == 'GP') &
    (ar['appointment_mode'] == 'Face-to-Face') &
    (ar['time_between_book_and_appointment'] == 'Unknown / Data Quality')
]

de1.sort_values('count_of_appointments')
```

[27]:

	icb_ons_code	appointment_month	appointment_status	hcp_type	appointment_mode	time_between_book_and_appointment	count_of_appointments
496290	E54000008	2020-01	Attended	GP	Face-to-Face	Unknown / Data Quality	1
505707	E54000008	2020-01	Attended	GP	Face-to-Face	Unknown / Data Quality	1
516594	E54000008	2020-01	Attended	GP	Face-to-Face	Unknown / Data Quality	1
487191	E54000008	2020-01	Attended	GP	Face-to-Face	Unknown / Data Quality	2
522453	E54000008	2020-01	Attended	GP	Face-to-Face	Unknown / Data Quality	2
491544	E54000008	2020-01	Attended	GP	Face-to-Face	Unknown / Data Quality	3
511261	E54000008	2020-01	Attended	GP	Face-to-Face	Unknown / Data Quality	3
529546	E54000008	2020-01	Attended	GP	Face-to-Face	Unknown / Data Quality	19

Figure 1 - Example of duplicated values.

When picking up an example from AR duplicated rows, it was noticed that it was not only a question of having some rows with the same values for all the columns but multiple entries for the same combination. In the datasets AD and NC, each row represents a unique combination of characteristics. Therefore, the decision made for AR was to do something similar. The multiple records for the same combination were considered incorrect data entries and something that would have to be checked with the

NHS team. It was essential to keep all the values, so a group of all the categorical columns was done with the sum of *count_of_appointments.*

The columns related to dates were also changed to reflect an adequate data type. An outliers check was done as well. The lower and upper limit values were calculated using the interquartile range method, and 46927 values for AR and 147958 for NC were found. A decision was made to keep them since the data had been cleaned previously, and the errors seemed to be important information that would completely alter the conclusions.

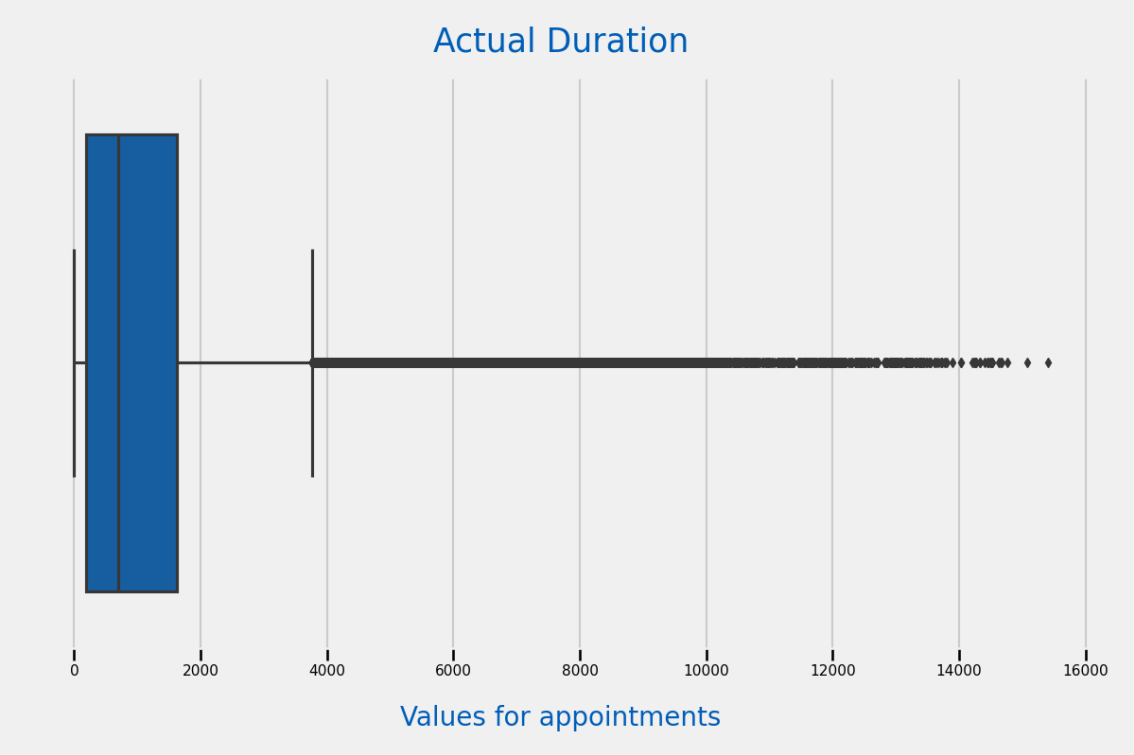

Figure 2 - Distribution of values of appointments for AD.

One problem with the data is that the sum of monthly appointments doesn't match all three datasets. The NHS team would have to be asked about this aspect. The AD and NC datasets (the ones that match) will be used as the reference for total appointments to overcome this issue.

Two more datasets were used, collected from external sources. The *nhs_regions* (more in Appendix A) is a categorical dataset that gives information about the regions, and it allowed us to build a reference dataset to connect all the others (more in Appendix B). The *nhs_staff* (more in Appendix C) shows the doctors for General Practices from December 2021 to June 2022.

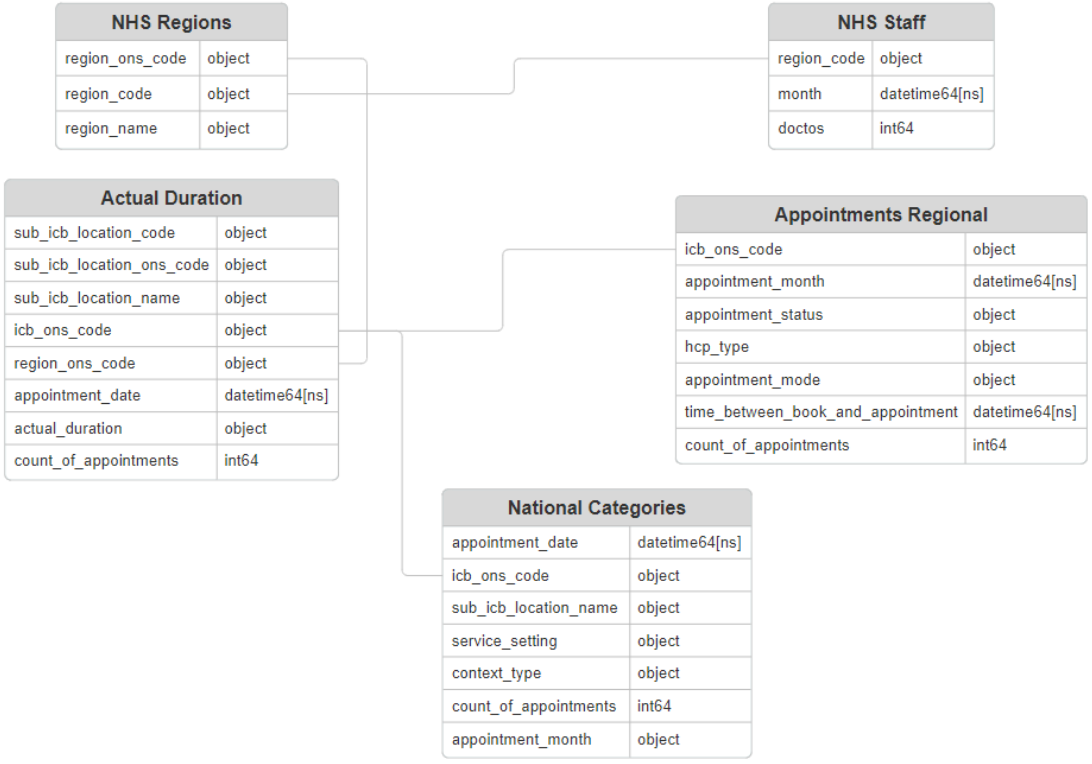

Figure 3 - Entity Relationship Diagram (ERP) of the initial data.

3. Visualisation and notebook structure

The Jupyter notebook was structured to have a clean look and to identify the section we are analysing quickly. The colour scheme is based on the NHS colour to be consistent with brand identity. That can be seen in the section titles of the notebook and on the graphs, mainly on the title, x-axis label, y-axis label, and some bar and line charts.

The analysis started with the univariate analysis, developing graphs for the categorical values to portray the most common characteristics of NHS according to the total appointments.

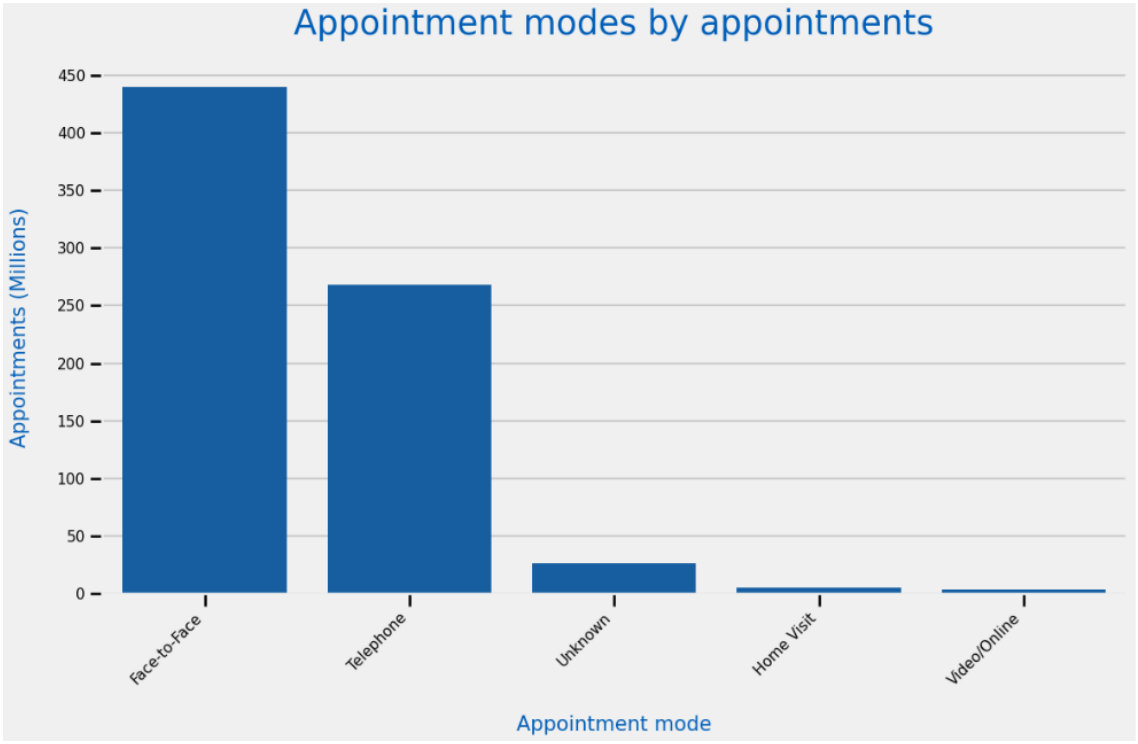

Figure 4 - Bar chart of the appointment modes.

The next step of the analysis was looking into the time evolution of some aspects of the data. In some variables, each characteristic's share was calculated to display its changes with the time evolution.

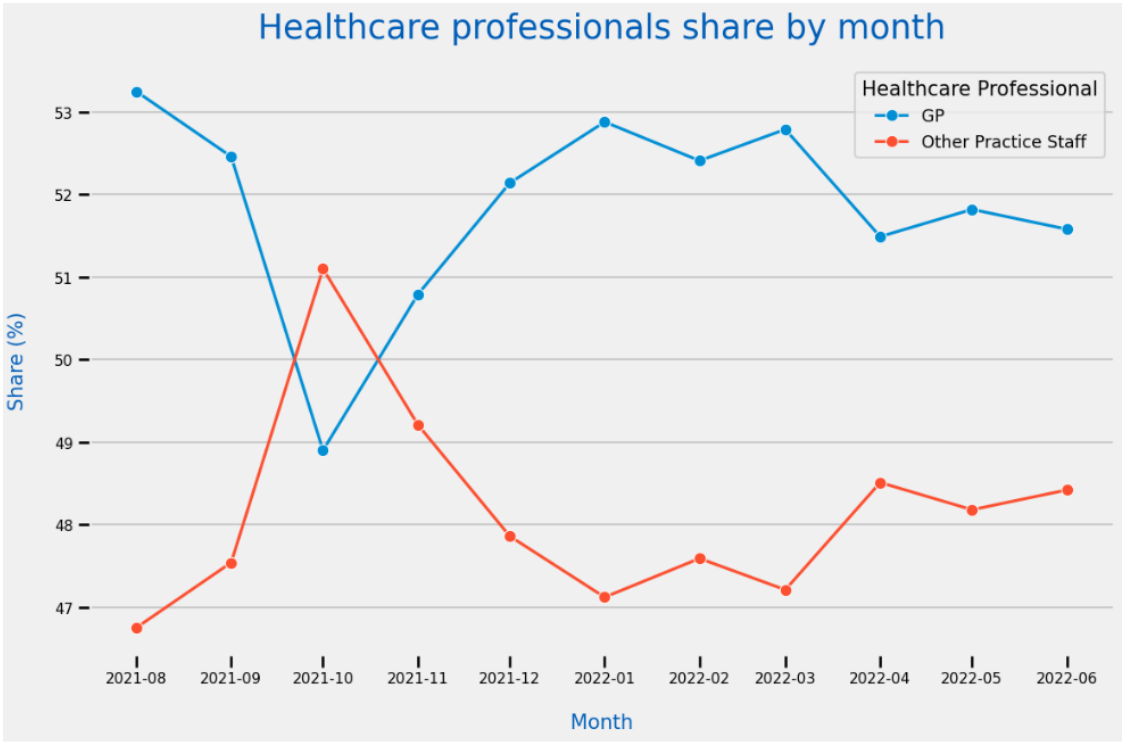

Figure 5 - Line chart with the shares of each healthcare professional type.

A social network, in this case, Twitter, was analyzed to provide extra material for insights. We leveraged Twitter for NHS data exploration, analyzing popular hashtags and planning future sentiment analysis to enhance engagement and communication strategies.

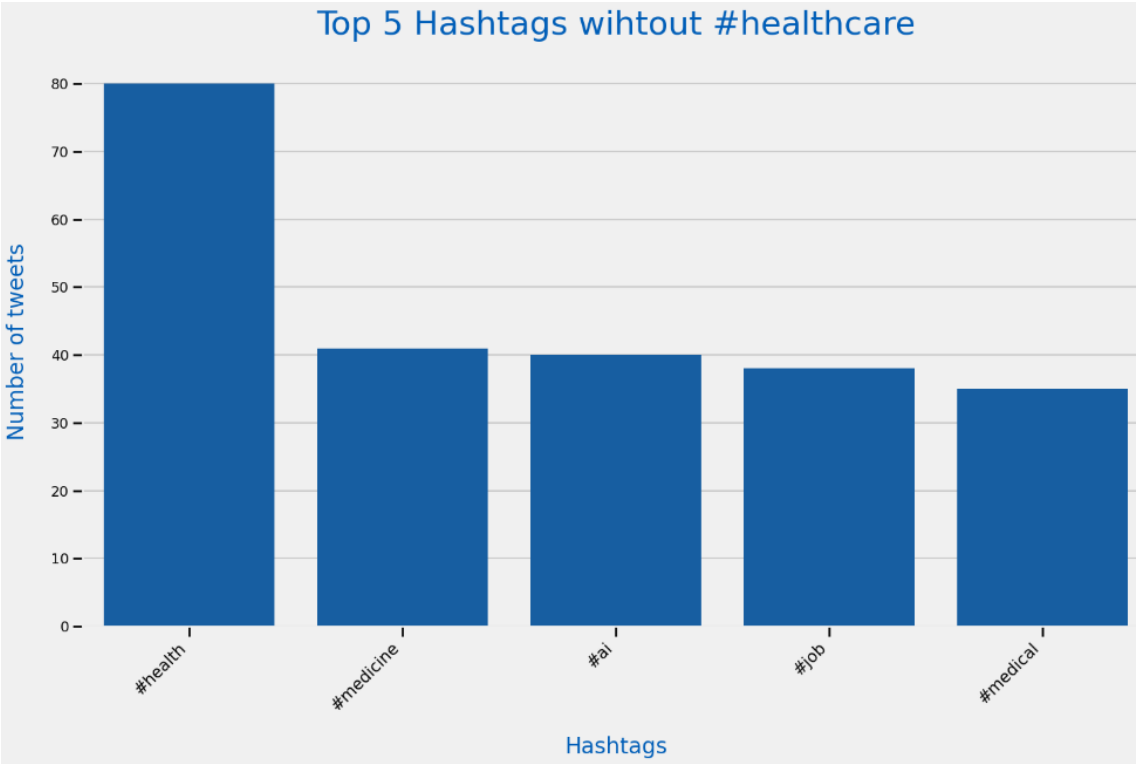

Figure 6 - Twitter hashtags analysis.

The last part of the report provided more profound insights into the business problem. The regional analysis identified differences between regions, helping to understand geographic disparities and resource allocation. The staff analysis assessed doctor workload by creating a metric of appointments per doctor, evaluating the burden on doctors overall and by region. Missed appointments were evaluated against region, mode, and time between booking and appointment to determine critical factors affecting appointment attendance. Finally, a daily capacity analysis compared actual appointments to NHS capacity guidelines, providing insights into NHS performance and potential overwhelm. By leveraging these comprehensive visualizations, we can enhance our strategic planning, improve resource management, and address critical operational challenges within the NHS.

4. Patterns, trends, and insights

Our analysis of the NHS data revealed crucial insights. We first noted a uniformity in the seasonality of total appointments per month across regions. However, a troubling trend emerged regarding the healthcare workforce. Over seven months, we observed a 1.91% decline in doctor numbers, amplifying workload pressures.

Further investigation unveiled significant regional disparities in appointment volumes, indicating an uneven distribution of medical staff and potential workload imbalances. These disparities were striking, with monthly appointment counts varying by as much as 169 appointments per doctor. Such discrepancies suggest that medical staff may not be evenly spread across regions, leading to localized areas experiencing higher workloads while others may have excess capacity.

Moreover, our analysis revealed a concerning uptick in missed appointments in recent years, contributing to rising NHS costs. Notably, we found that appointments booked for the same day had a remarkably low missed appointment rate of less than 2%, while appointments booked over 28 days in advance had a missed appointment rate of over 10%.

Lastly, our analysis highlighted persistent overcapacity issues, with NHS facilities consistently exceeding the guideline maximum of 1.2 million appointments per day, particularly on weekdays.

Moving forward, exploring different staff types, assessing resource capacity, and analysing regional appointment rates relative to population density are crucial steps. These insights will enable the NHS to make informed decisions and optimize service delivery for the benefit of all stakeholders.

Appendices

Appendix A – Business questions

The business questions clarify objectives, guide data collection, and focus analysis, ensuring solutions are directly aligned with addressing specific business problems effectively and efficiently.

Staff Capacity

- How has staff capacity evolved over the months?
- What are the differences in staff capacity between the regions?

Missed Appointments

- What is the rate of missed appointments?
- Are there any patterns in missed appointments?

Resources Capacity

- Is NHS under pressure and working above the recommended capacity?

Appendix B – NHS Regions

This data frame was collected from the Office for National Statistics[1] and represents the NHS England region names and codes as of July 2022.

	NHSER22CD	NHSER22CDH	NHSER22NM
0	E40000003	Y56	London
1	E40000005	Y59	South East
2	E40000006	Y58	South West
3	E40000007	Y61	East of England
4	E40000010	Y62	North West
5	E40000011	Y60	Midlands
6	E40000012	Y63	North East and Yorkshire

Figure 7 - NHS Regions.

[1] Link - https://geoportal.statistics.gov.uk/documents/46b634b42ceb45cbbfbe9c960fb77ec9/about

Appendix C – Regions reference data frame

The NHS regions collected were used to create a region reference data frame, which is essential to connect several data frames using the *icb_ons_code* available in the three data frames provided by the NHS.

The steps to create the data frame can be viewed in the Jupyter Notebook. Each *icb_ons_code* is unique.

	icb_ons_code	region_ons_code	region_code	region_name
0	E54000050	E40000012	Y63	North East and Yorkshire
1	E54000048	E40000010	Y62	North West
2	E54000057	E40000010	Y62	North West
3	E54000008	E40000010	Y62	North West
4	E54000061	E40000012	Y63	North East and Yorkshire
5	E54000060	E40000011	Y60	Midlands
6	E54000054	E40000012	Y63	North East and Yorkshire
7	E54000051	E40000012	Y63	North East and Yorkshire
8	E54000015	E40000011	Y60	Midlands
9	E54000010	E40000011	Y60	Midlands

Figure 8 - First rows of Regions data frame.

Appendix D – NHS Staff

This data frame was collected from the NHS Digital statistics[2] website. From December 2021 onwards, the NHS started making monthly data available, including the number of general practitioner doctors and several other staff members. The regions and sub-regions in the files match the ones we have in the datasets. The number of GPs in full-time equivalent per region was taken for this analysis. A future step in this work would be analysing the staff according to different types, not only doctors.

England		All GPs	GP Partners	Salaried GPs	GPs in Training Grade	GP Retainers	GP Regular Locums
England		36,009	16,937	9,803	8,252	245	772
Y56	London	5,334	2,384	1,877	836	23	213
Y58	South West	3,704	1,871	1,086	630	55	61
Y59	South East	4,956	2,570	1,461	792	60	73
Y60	Midlands	7,050	3,301	1,732	1,815	37	166
Y61	East of England	3,779	1,968	959	740	19	94
Y62	North West	4,828	2,180	1,232	1,291	20	105
Y63	North East and Yorkshire	5,437	2,663	1,455	1,227	31	60

Figure 9 - GWP Bulletin Tables - January 2022. One example of the data taken in the file for January 2022.

	NHS England Region Code	2021-12	2022-01	2022-02	2022-03	2022-04	2022-05	2022-06
0	Y56	5384	5334	5356	5336	5333	5287	5282
1	Y58	3731	3704	3744	3721	3709	3679	3649
2	Y59	4984	4956	4963	4938	4937	4950	4930
3	Y60	7055	7050	6989	6960	6945	6871	6923
4	Y61	3794	3779	3810	3794	3785	3768	3738
5	Y62	4823	4828	4893	4879	4802	4777	4740
6	Y63	5487	5437	5507	5486	5461	5428	5324

Figure 10 - NHS Staff.

[2] Link - https://digital.nhs.uk/data-and-information/publications/statistical/general-and-personal-medical-services