NHS DATA PROJECT

Assignment

Course 2: Data Analytics using Python

LSE Data Analytics Online Career Accelerator

Tamas Balog 24 October 2022

Word limit 1000

At first I had to understand the data structure and size, I focused on extracting information on timeframe and appointment count and characteristics.

I created and validated dataframes, explored metadata. I created an excel file to manage basic information about the data set and the files, later I added more information to this table.

actual duration		
(137793, 8)		
sub_icb_location_code	object	loc
sub_icb_location_ons_code	object	loc
sub_icb_location_name	object	loc
icb_ons_code	object	loc
region_ons_code	object	loc
appointment_date	object	time
actual_duration	object	time
count_of_appointments	int64	value
total number of appointments		167,980,692
records 137793		
ons		42
sub icb		106
start of period		Dec-21
end of period		Jun-22
period length		7 month
count_of_appointments		
count		137793
mean		1219.080011
std		1546.902956
min		1
0.25		194
0.5		696
0.75		1621
max		15400

appointment regional		
(596821, 7)		
icb_ons_code	object	loc
appointment_month	object	month
appointment_status	object	status
hcp_type	object	type
appointment_mode	object	type
time_between_book_and_appointment	object	time
count_of_appointments	int64	value
total number of appointments		742,804,525
records 596821		
ons		42
start of period		Jan-20
end of period		Jun-22
period length		28 month
count_of_appointments		
count		596821
mean		1244.601857
std		5856.887042
min		1
0.25		7
0.5		47
0.75		308
max		211265

national categories		
(817394, 8)		
appointment_date	datetime64[ns]	time
icb_ons_code	object	loc
sub_icb_location_name	object	loc
service_setting	object	type
context_type	object	type
national_category	object	type
count_of_appointments	int64	value
appointment_month	datetime64[ns]	time
total number of appointments		296,046,770
records 817394		
ons		42
sub icb		106
start of period		Aug-21
end of period		Jun-22
period length		11 month
count_of_appointments		
count		817394
mean		362.183684
std		1084.5766
min		1
0.25		7
0.5		25
0.75		128
max		16590

Once I have some knowledge of the data I reviewed the context and tried to Identify key objectives and metrics, I summarized the objectives of the analysis into the starting slide of my video presentation. I found that the focus from stakeholders is on capacity, utilisation and most of all cost of non-attendance for appointments.

From the start I kept asking questions both about the data and the objectives and tried to explore more and more.

After completing the three questions for Week2 assignment:

1. There are 106 locations in 42 groups in the dataset.
2. The locations with the highest number of records
   NHS Norfolk and Waveney ICB - 26A
   NHS Kent and Medway ICB - 91Q
   NHS North West London ICB - W2U3Z
   NHS Bedfordshire Luton and Milton Keynes ICB - M1J4Y
   NHS Greater Manchester ICB - 14L
3. Appointment types:
   service settings: 5
   context types: 3
   national categories: 18
   appointment statuses: 3

Additional checks:

The number of appointments grouped by actual duration:

show the large percentage 24% of unknown appointment length. It also indicates that most appointments are shorter than 15 minutes.For me this is useful information, stakeholders might be familiar with these facts, there is communication needed about that, so I included this in video presentation.

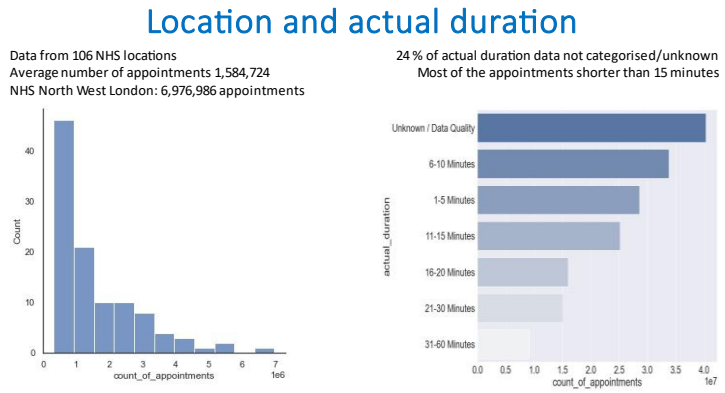
Location:

104 locations NHS North West London has the highest number of appointments 6,976,986 vs average 1,584,724 appointments for all locations

Timeframe:

overall the records in the ad file are from 01-12-2021 – 30-06- 2022

I added two visuals:



Further analysis needed:

- This is a national data, perhaps it includes regional factors. These can be identified one way. The other way would be to examine one location and make a more detailed analysis of that.

Appointment status:

This is the key aspect for the analysis, it shows up in the appointment regional data. I identified the attendance ratio metric and created a visual to show it over time

Created a lineplot showing hcp_type over time.

Created a lineplot showing appointment mode over time.

The Face-to-Face and Telephone appointments both fluctuate. These two appointment modes together became the main channels from summer 2020.

Further analysis can be done to find out if any of these two methods have different attendance ratio.

Increasing Video/Online appointments could be a way to efficiently use resources

Further analysis needed :

how the changes in the two main channels is related to change in total number of appointments? Why and if there is little impact on other categories at the same time?

analyse relationship between appointment status and booking time, appointment mode, hcp type. Perhaps compare regional data.

This stage of the processing was really an iterative work with continuously developing insight.

For the visuals I used lineplots that are suitable to show time-series.

Then, for assignment week 3 I have completed these checks:

1. Dataframes cover these dates
   ar: 2020-01 to 2022-06,
   ad: 2021-12-01 to 2022-06-30,
   nc: 2021-08-01 to 2022-06-30

The presentation will explain this to stakeholders

2. Service settings reported in NW London in the specified period:

General Practice	4804239
Unmapped	391106
Other	152897
Primary Care Network	109840
Extended Access Provision	9815
total	5556241

3. Highest number of appointments 2021-11 it is 30,405,070. This is relevant for the analysis and I included this data into the presentation.

4. Total number of records per month:

Month	Number of records
(2021, 8)	69999
(2021, 9)	74922
(2021, 10)	74078
(2021, 11)	77652
(2021, 12)	72651
(2022, 1)	71896
(2022, 2)	71769
(2022, 3)	82822
(2022, 4)	70012
(2022, 5)	77425
(2022, 6)	74168
total	817394

Additional checks:

Further analysis needed:

> Investigate appointments booked on the same day, relationship with attended status, are these walk-in appointments?

Then, for assignment Week 4

I created three visualisations indicating the number of appointments per month for service settings, context types, and national categories:

> One category is dominant in each visual: general practice, care related encounter, general consultation routine, need to remove those from visuals to see patterns for others
>
> Looking at daily visualisation of service settings shows more volatility with daily and weekly changes.

I created four visualisations indicating the number of appointments for service setting per season:

> Seasonal visuals are different, highest number of appointments in autumn, further weekly patterns can be identified, highest number of appointments seem to be on the start of the week
>
> Perhaps these patterns could be investigated further in relation of attendance and capacity.

Then, for assignment Week 5

Value counts of favourites and retweets show that only some tweets are really successful. It would make sense focusing on those. However, when extracting data from social media it also makes sense to consider who are the people who are most active, are we focusing on celebrities and influencers, is this a limited group, are they or the retweets genuine?

I have changed the colour palette and removed the healthcare and health hashtags, this will need to be explained to the stakeholders. I also removed hashtags with less than 12 counts to focus on the trending subjects.

Covid came up in different hashtags so that has to be considered and communicated.

Then, for assignment Week 6

I calculated utilisation and created a visual

Visualised HCP types, appointment status over time there, I see no significant trend other than the dominant categories

Visualising appintment_mode shows that in the busiest months there is a significant rise in face-to-face appointments

investigate further:

> The period with highest number of appointments 2021-11, so it would make sense looking into the details which trends changed, which regional and other features were most stressed. That period could be examined broken down to daily data

Summary:

Should the NHS start looking at increasing staff levels? Based on the visualisations there is no straight forward answer. There are dangerous peaks of utilisation with General Practice remaining the key service setting. Exploring flexible resources and ways to do and book appointments would be the best. Changing to resource efficient channels like online and video appointments could improve attendance issues and save costs. The data shows significant drops in attendance in certain periods this needs to be investigated further. The fact that In the busiest months there is a significant rise in face-to-face appointments also suggests alternative ways of appointments are needed.

Word count 1090