Course 2 Assignment

Diagnostic Analysis Using Python

Damian Ferguson

2nd March 2024

Background

The NHS needs the right capacity to support an increasing population. Some stakeholders believe capacity should be added through investment others that resources should be better utilised. The NHS needs to understand current utilisation and trends within its' network to make the right decision(s).

The NHS believes missed GP appointments contribute to lower utilisation. To determine if this is true, utilisation of current resources needs to be understood and two questions have been posed:

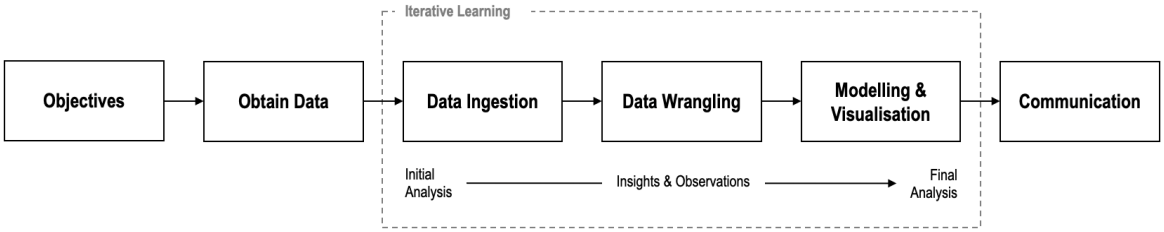
- Has there been adequate staff and capacity?
- What was the actual resource utilisation?

Focus:

A. Is there adequate staff and capacity in the network (utilisation)?
B. Do missed GP appointments cause low utilisation?

Analytical approach

The following approach was used:



Data Ingestion

The three main data sets were imported into dataframes. Field names, data types and null records were checked. Dataframe heads were inspected; descriptive statistics viewed; and the sum of appointments calculated.

After inspection, integers were changed to float; and date objects converted to datetime. New column '*appointment_month*' was added to '*ad*' Dataframe date ranges were determined. Sum of appointments for overlapping dates were also determined.

Dataframes and column names were reviewed for commonality (Appendix A). Value counts were conducted on categorical variables (Appendix B). Columns relating to a '*location*' or '*ONS*' code were reviewed for redundancy (Appendix C).

Summary of Data Ingestion	Data Frames		
	ad	ar	nc
File Name	actual_duration	appointments_regional	national_categories
File Type	.csv	.csv	.xlsx
Columns x Rows	8 x 137,793	7 x 596,821	8 x 817,394
Missing Values	None	None	None
'count_of_appointments' converted 'int' to 'float' (for arithmetic operations)	✓	✓	✓
'appointment_date' converted 'object' to 'date' (for datetime operations)	✓	✓	Column already in date format
'appointment_month' converted 'object' to 'period' (for datetime operations)	New column added, derived from 'appointment_date'	✓	✓
Date Range Start	01/12/2021	2020-01	01/08/2021
Date Range End	30/06/2022	2022-06	30/06/2022
Sum of 'count_of_appointments'	167,980,692	742,804,525	296,046,770
Sum of 'count_of_appointments' where dates overlap.	167,980,692	182,963,194	182,963,194

Observations:

- '*count_of_appointments*', '*icb_ons_code*' and '*appointment_month*' are common across dataframes
- *icb_ons_code*' and '*appointment_month*' are potential join keys.
- Data only overlaps between December 2021 and June 2022.
- Sum of appointments varies between dataframes but is related to date range.
- During overlapping dates, 'ar' and 'nc' have the same sum of appointments (may be from the same data source).

Decisions:

- '*count_of_appointments*' will be the focus of analysis.
- Analysis will be performed in monthly buckets, '*appointment_month*' being a common column.
- Quoted capacity is 1.2M appointments per day, 36M per month.
- '*icb_ons_code*' is the only common (useful) '*location'* or '*ONS'* column therefore '*region_ons_code*', '*sub_icb_location_ons_code*', '*sub_icb_location_code*' and '*sub_icb_location_name*' will be deleted.
- Other than looking for outliers, no further data cleansing will be done.

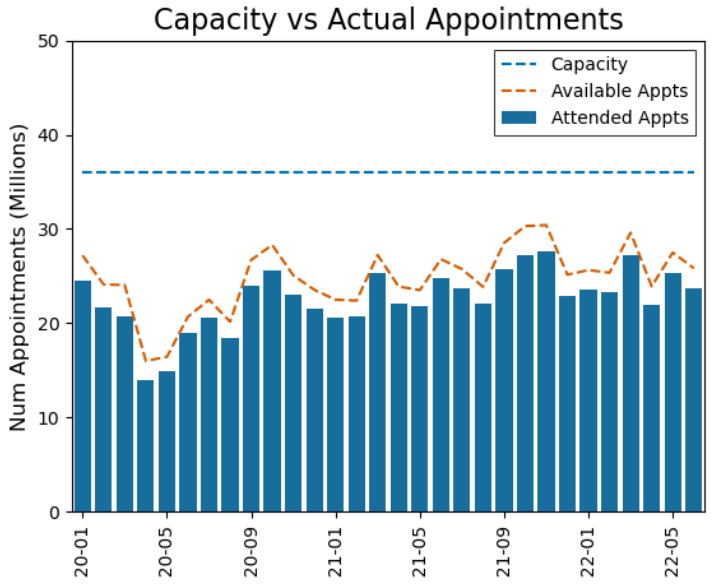
Data Wrangling

Two data subsets were initially created with further subsetting during visualization and modelling. '***utilisation***' was used to summarize appointments by month and calculate utilization percentages (Appendix D). '***missed***' was used hold a simple subset of missed appointments data (Appendix E). Both were derived from 'ar' dataframe.

Modelling & Visualisation

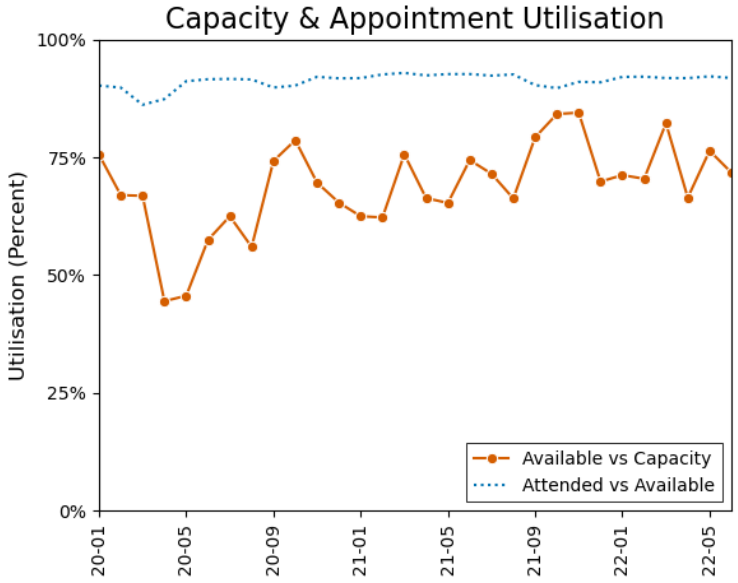
Is there adequate staff and capacity in the network (utilisation)?

The number of available appointments is significantly lower than quoted capacity, whereas attendance is close to the number of available appointments.



There were two outliers (April and May 2020) which occurred during COVID lockdown. As they appear to have no material effect, they were not removed.

Capacity and appointment utilization were calculated (Appendix F).



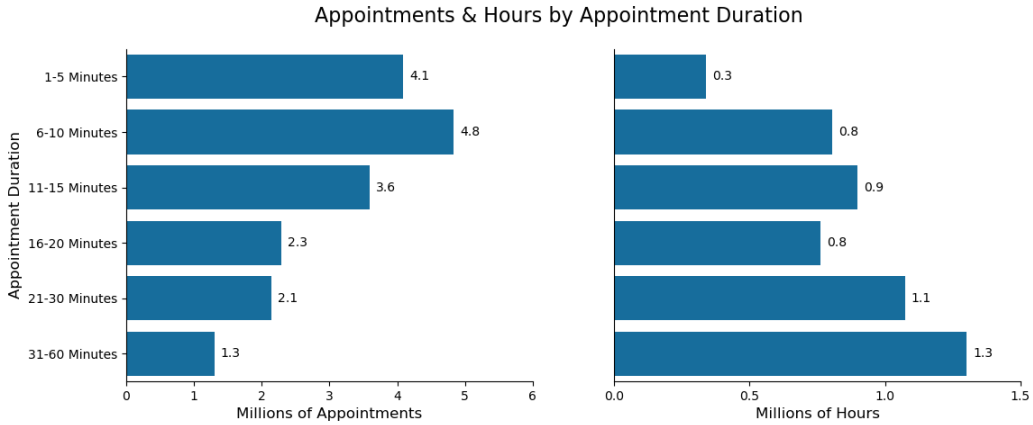
Capacity utilization ranges between 44% and 75% (mean = 68%); attendance ranges between 86% and 92% (mean = 91%). Capacity utilization seems low.

The number of available appointments is assumed to be a function of staff and appointment mix. Staffing data was unavailable therefore, appointment mix became the focus. This was in the '*ad'* dataframe.

Source data only held number of appointments. A new subset (*'ad1_subset1_appt_hrs')* was created for hours (Appendix F). Hours were calculated based on the maximum duration of the categorised duration multiplied by the number of appointments.

There are only 7 months of data, therefore no conclusions could be drawn about impact over time (Appendix G).

The number of appointments and hours per month were plotted side by side to visualise the relationship between them.



31-60 min appointments account for 7% of appointments and 25% of time. 21-30 min appointments account for 12% of appointments and 21% of time. In total, 19% of appointments take up 46% of the available time.

An Excel model was created to view the potential impact of reducing the number of long appointments. It shows that a 10% reduction in 21-30 and 31-60 minute appointments would create 7.2M additional 11-15 minute appointments.

Appt Duration	Appts per Month	10% Appts	Equiv 11-15 Min Appts	Net Monthly Impact	Annual Impact
21-30 Minutes	2.1	0.21	0.42	0.21	2.52
31-60 Minutes	1.3	0.13	0.52	0.39	4.68
Totals	**3.4**	**0.34**	**0.94**	**0.6**	**7.2**

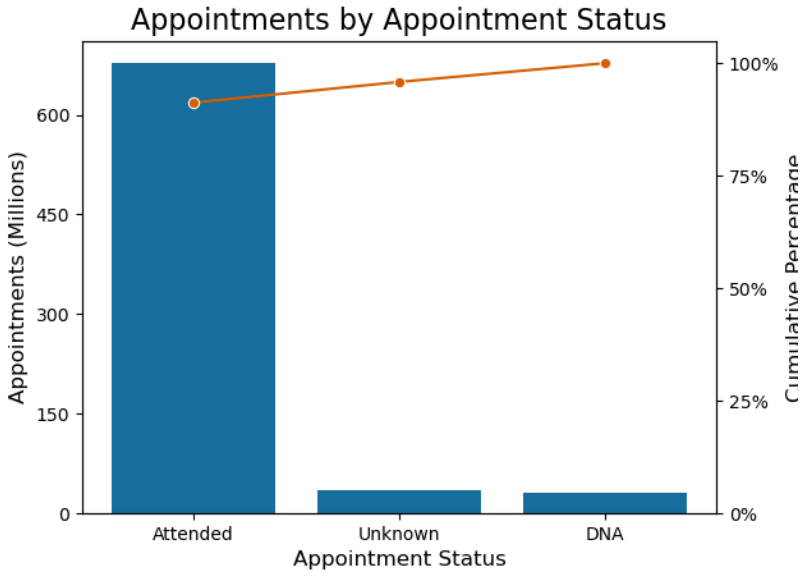
Based on 742.8M appointments, this would equate to a 1% increase overall

This alone will not close the capacity gap and suggests that staffing levels should be reviewed.

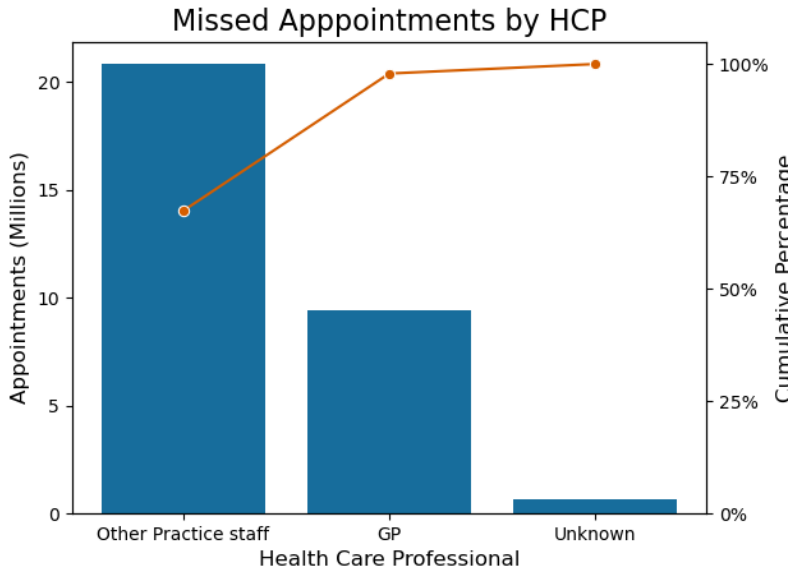
24% of appointments were observed with an unknown duration (Appendix G). This should be reviewed.

Do missed GP appointments cause low resource utilisation?

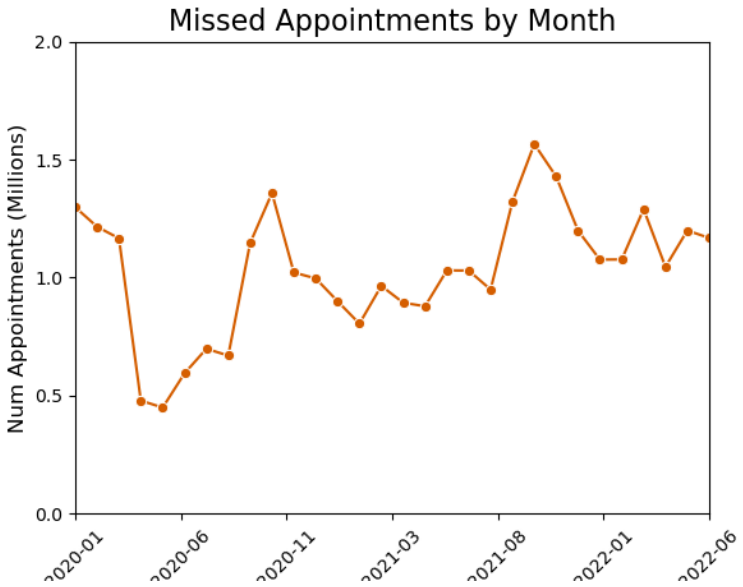
Attendance of appointments ranges between 86% and 92% (mean = 91%). 4.3% of appointments are missed (DNA), 4.7% are of unknown status.



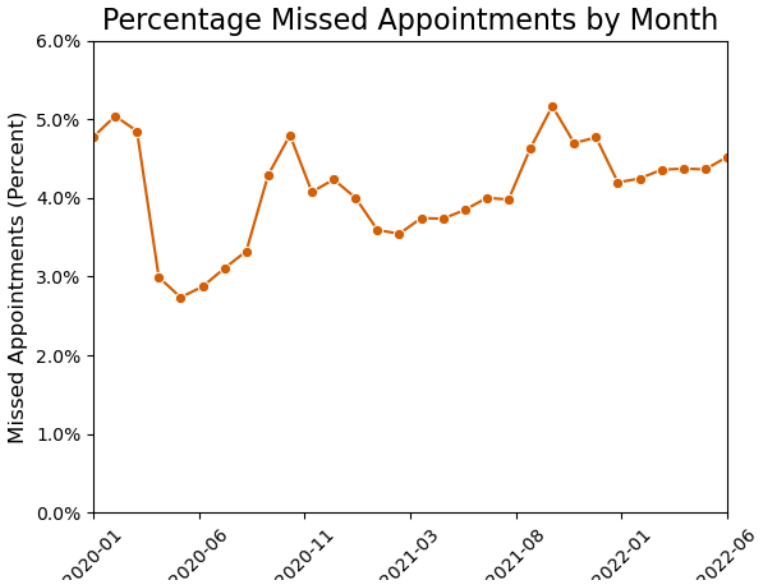
65% of missed appointments are with Other Practice Staff and 30% with GPs.



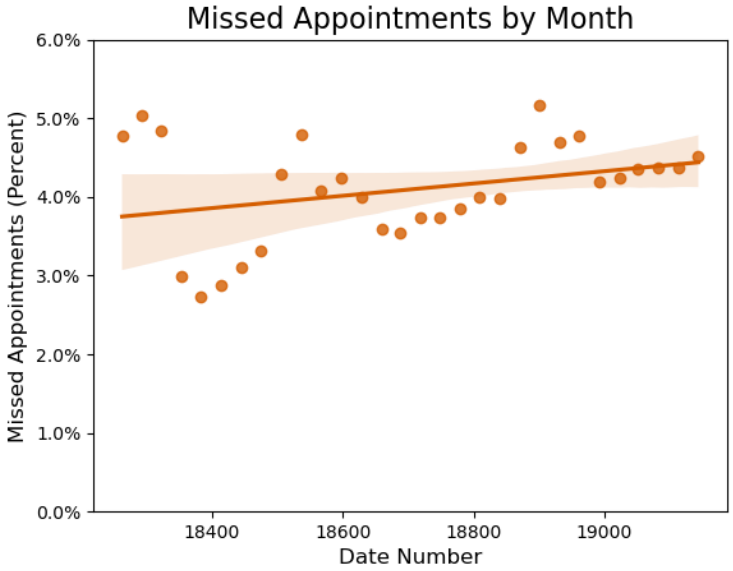
Missed appointments are increasing over time and are currently in the range of 1.0M to 1.5M per month.



The percentage of missed appointments is increasing over time too.



This trend can be seen better in the chart below rising 0.6% over 30 months. If this trend continues there will be a further increase of 0.24% (0.84M) in the next twelve months.



Although this is concerning it is insignificant compared to capacity utilisation discussed earlier.

National categories data

During the current analysis, this data was not used. It could be used for future analysis.

Tweet Data

A separate request to determine if X / twitter data would augment the main analysis. After a brief review, the data was found to contain three references to the NHS, none relating to main issue (Appendix E). Consequently, this investigation was halted.

Conclusions & Recommendations

Is there adequate staff and capacity?

No. Available appointments range between 44% and 75% of quoted capacity. Reducing or limiting the number of appointments over 20 mins by 10% would result in an annual increase in available appointments of 1% or 7.2M (Priority # 2). This alone will not close the capacity gap. Staffing levels will need to be reviewed (Priority #1).

Do missed GP appointments cause low utilisation?

Not significantly. Missed appointments with Other Practice Staff accounts for 65% of missed appointments, GPs 30%. Although reducing this would be helpful, it would not materially impact on capacity utilization overall (Priority # 3).

Further analysis and investigation

A review of quoted capacity should be conducted to make sure it is aligned to patient demand. Adding unneeded capacity would waste money and resources. Not adding when needed would impact patient care.

Further investigation should be carried out to understand and resolve unknown data values, 4.7% of appointment status and 24% of appointment duration.

APPENDIX

A. Common and unique dataframe columns

	Data Frame		
Field Name	**ad**	**ar**	**nc**
actual_duration	✓		
appointment_date	✓		✓
appointment_mode		✓	
appointment_month	✓	✓	✓
appointment_status		✓	
context_type			✓
count_of_appointments	✓	✓	✓
hcp_type		✓	
icb_ons_code	✓	✓	✓
national_category			✓
region_ons_code	✓		
service_setting			✓
sub_icb_location_code	✓		
sub_icb_location_name	✓		✓
sub_icb_location_ons_code	✓		
time_between_book_and_appointment		✓	

B. Categorical variable value counts

Field Name	Unique Values
actual_duration	7
appointment_mode	5
appointment_status	3
context_type	3
hcp_type	3
icb_ons_code	42
national_category	18
region_ons_code	7
service_setting	5
sub_icb_location_code	106
sub_icb_location_name	106
sub_icb_location_ons_code	106
time_between_book_and_appointment	8

C. Understanding ‘location’ and ‘ONS’ related columns

Understanding 'location' and 'ONS' columns

42 unique values, represents an **Integrated Care Board** ; common to

region_ons_code	icb_ons_code	sub_icb_location_ons_code	sub_icb_location_code	sub_icb_location_name
E40000010	E54000048	E38000034	00X	NHS Lancashire and South Cumbria ICB - 00X
E40000010	E54000048	E38000050	01A	NHS Lancashire and South Cumbria ICB - 01A
E40000012	E54000061	E38000044	02X	NHS South Yorkshire ICB - 02X
E40000010	E54000008	E38000194	02E	NHS Cheshire and Merseyside ICB - 02E
E40000011	E54000062	E38000259	D2P2L	NHS Black Country ICB - D2P2L

7 values, represents a **Region** , highest level of the hierarchy; unique to 'ad' dataframe

106 unique values in each column; represents a **Location** within ICB; location code embedded within location name; unique to 'ad' dataframe

Relationships:

Each reagion has 1 or many ICBs (1:M)
Each ICB has 1 or many locations (1:M)
Location ONS code, location code and location name all have a 1 to relatiomship (1:1)

D. Utilisation data subset

```
# Create a subset of appointments data by appointment status.
utilisation = appts_subset.pivot(index = 'appointment_month', columns = 'appointment_status',
                                 values = 'appts_M')

# Add a total value.
utilisation['total_available_appts'] = utilisation['Attended'] + utilisation['DNA'] + utilisation['Unknown']

# Add capcity column based on 1.2M appts and 30 days per month.
for each in utilisation:
    utilisation['capacity'] = (1.2 * 30)

# Add capacity utilisation columns - one based on quoted capcity figure, on based on total available appointments.
utilisation['capacity_util'] = (utilisation['Attended'] / utilisation['capacity']) * 100
utilisation['available_util'] = (utilisation['Attended'] / utilisation['total_available_appts']) * 100
utilisation['available_vs_capacity'] = (utilisation['total_available_appts'] / utilisation['capacity']) * 100
```

appointment_status	appointment_month	attended	missed	unknown	total_available_appts	capacity	capacity_util	available_util	available_vs_capacity
0	2020-01	24.538291	1.298269	1.362736	27.199296	36.0	68.161919	90.216640	75.553600
1	2020-02	21.640067	1.215154	1.249400	24.104621	36.0	60.111297	89.775595	66.957281
2	2020-03	20.718865	1.166314	2.168289	24.053468	36.0	57.552403	86.136706	66.815189
3	2020-04	13.982824	0.478766	1.546291	16.007881	36.0	38.841178	87.349625	44.466336
4	2020-05	14.962850	0.449057	1.005305	16.417212	36.0	41.563472	91.141236	45.603367

E. Missed data subset

```
# Create a new dataframe for missed appointments only.
missed = pd.DataFrame(data = ar1) [(ar1['appointment_status'] == 'DNA')]

# Convert date to str for analysis purposes.
missed['appointment_month'] = missed['appointment_month'].astype(str)
```

	icb_ons_code	appointment_month	appointment_status	hcp_type	appointment_mode	time_between_book_and_appointment	count_of_appointments
43112	E54000051	2021-09	DNA	GP	Face-to-Face	More than 28 Days	53.0
49203	E54000051	2022-01	DNA	Unknown	Telephone	Same Day	10.0
125986	E54000054	2021-10	DNA	GP	Home Visit	Same Day	1.0
361569	E54000026	2022-06	DNA	GP	Face-to-Face	Same Day	116.0
203042	E54000057	2021-04	DNA	Other Practice staff	Unknown	15 to 21 Days	121.0

F. Utilization calculations

Utilization Calculations				
Capacity Utilization	=	$\frac{\text{Total Available Appointments}}{\text{Capcity}}$	x 100	Are enough appoints being provided?
Appointment Utilization	=	$\frac{\text{Attended Appointments}}{\text{Total Available Appointments}}$	x 100	Are appointments being attended?

G. ad1_subset1_appts_hrs data subset

```
# Pivot ad1 subset 1 to get rows into columns for appointment status.
ad1_subset1_appts_pvt = ad1_subset1_appts.pivot(index = 'appointment_month', columns = 'actual_duration',
                                values = 'count_of_appointments').reset_index()

ad1_subset1_appts_pvt['appointment_month'] = ad1_subset1_appts_pvt['appointment_month'].astype(str)

ad1_subset1_appts_pvt
```

actual_duration	appointment_month	1-5 Minutes	11-15 Minutes	16-20 Minutes	21-30 Minutes	31-60 Minutes	6-10 Minutes
0	2021-12	4.266686	3.280132	2.023774	1.885864	1.163515	4.654000
1	2022-01	3.975252	3.537308	2.242739	2.125583	1.310454	4.778719
2	2022-02	3.908364	3.508227	2.238822	2.116476	1.295696	4.695914
3	2022-03	4.570114	4.095290	2.607518	2.450292	1.485532	5.489184
4	2022-04	3.714437	3.297961	2.095851	1.958618	1.177878	4.422913
5	2022-05	4.203478	3.845162	2.481613	2.330326	1.390177	5.033645
6	2022-06	3.962534	3.596802	2.313930	2.159206	1.280180	4.726440

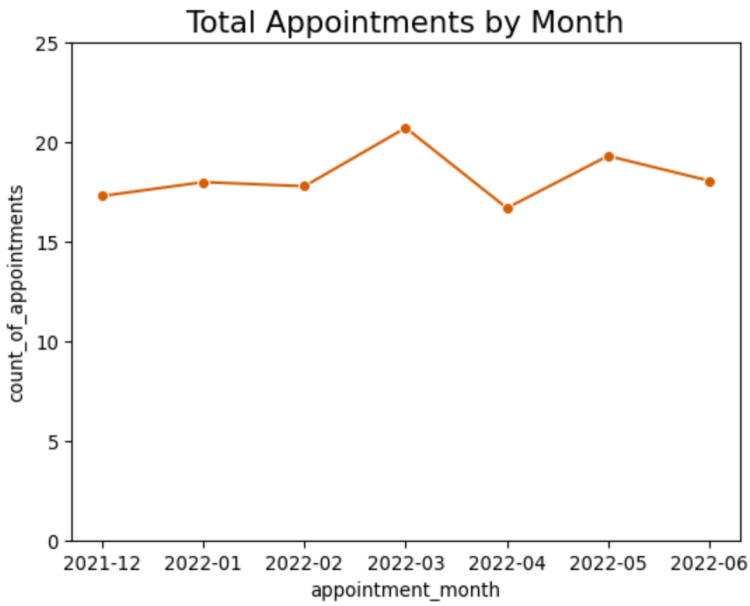
```
# Calculate total timee for each column num appts (Ms) *  mins / 60 in M hrs.
# Assume appts take the max time in the range.
ad1_subset1_appts_pvt['1-5 Minutes'] = (ad1_subset1_appts_pvt['1-5 Minutes'] * 5) / 60
ad1_subset1_appts_pvt['11-15 Minutes'] = (ad1_subset1_appts_pvt['11-15 Minutes'] * 15) / 60
ad1_subset1_appts_pvt['16-20 Minutes'] = (ad1_subset1_appts_pvt['16-20 Minutes'] * 20) / 60
ad1_subset1_appts_pvt['21-30 Minutes'] = (ad1_subset1_appts_pvt['21-30 Minutes'] * 30) / 60
ad1_subset1_appts_pvt['31-60 Minutes'] = (ad1_subset1_appts_pvt['31-60 Minutes'] * 60) / 60
ad1_subset1_appts_pvt['6-10 Minutes'] = (ad1_subset1_appts_pvt['6-10 Minutes'] * 10) / 60

ad1_subset1_appts_pvt
```

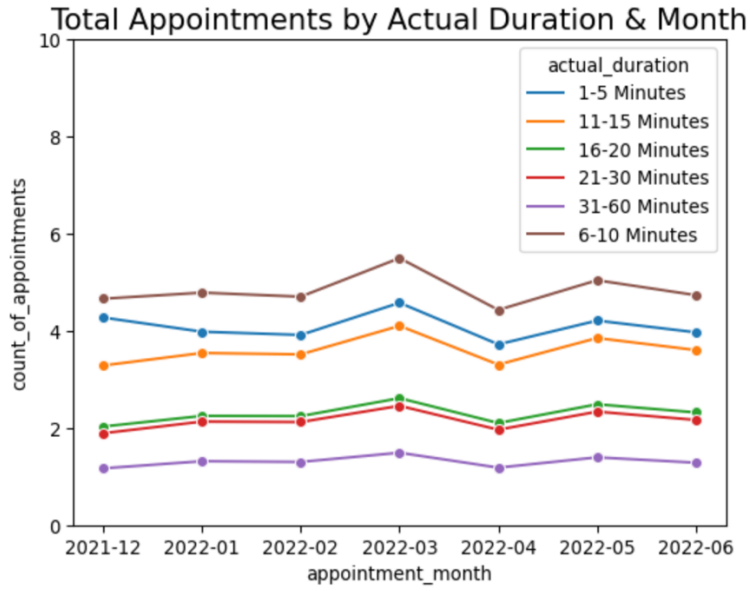
actual_duration	appointment_month	1-5 Minutes	11-15 Minutes	16-20 Minutes	21-30 Minutes	31-60 Minutes	6-10 Minutes
0	2021-12	0.355557	0.820033	0.674591	0.942932	1.163515	0.775667
1	2022-01	0.331271	0.884327	0.747580	1.062791	1.310454	0.796453
2	2022-02	0.325697	0.877057	0.746274	1.058238	1.295696	0.782652
3	2022-03	0.380843	1.023823	0.869173	1.225146	1.485532	0.914864
4	2022-04	0.309536	0.824490	0.698617	0.979309	1.177878	0.737152
5	2022-05	0.350290	0.961291	0.827204	1.165163	1.390177	0.838941

```
# Re-shape the data and change value column name.
ad1_subset1_hrs = ad1_subset1_appts_pvt.melt(id_vars = ['appointment_month'],
                                value_vars = ['1-5 Minutes', '11-15 Minutes',
                                              '16-20 Minutes', '21-30 Minutes',
                                              '31-60 Minutes', '6-10 Minutes'])

ad1_subset1_hrs.rename(columns = {'value': 'hours_Ms'}, inplace = True)

ad1_subset1_hrs.head()
```

	appointment_month	actual_duration	hours_Ms
0	2021-12	1-5 Minutes	0.355557
1	2022-01	1-5 Minutes	0.331271
2	2022-02	1-5 Minutes	0.325697
3	2022-03	1-5 Minutes	0.380843
4	2022-04	1-5 Minutes	0.309536

H. ad1_subset1_appts_hrs data subset





I. Unknown appointment durations