
# DIAGNOSTIC ANALYSIS
## (USING PYTHON)

# TABLE OF CONTENTS

***"Wherever the art of Medicine is loved, there is also a love of Humanity."— Hippocrates***



| | |
|---|---|
| SUBMITTED BY | Smita Prasad |
| COURSE | COURSE 2 DATA ANALYTICS USING PYTHON |
| DOCUMENT LENGTH | 1085 WORDS (EXCLUDING COVER) |

# LSE DA ACCELERATOR ANALYSIS USING PYTHON

## CONTEXT

A team of data analysts was contracted by the National Health Services (NHS), a publicly funded healthcare system in England to analyze the data they collect to adopt a more data-informed approach in deciding how best to handle their problems. The aim is to get started with the data exploration, data wrangling, visualizations, and identifying possible trends in the data sets provided to gain meaningful insights that can inform decision making.

## APPROACH

The dataset includes the following datasets-:

1. actual_duration.csv – Details of appointments made by patients. For example, the regional information, date, duration, and count of appointments.
2. appointments_regional.csv – Details on the type of appointments made by patients. For example, regional information, the month of appointment, appointment status, healthcare professional, appointment mode, the time between booking and the appointment, the count of appointments.
3. national_categories.xlsx – Details of the national categories of appointments made by patients. For example, the regional information, date of appointment, service setting, type of context, national category, and the number of appointments pertaining to a certain class.
4. metadata_nhs.txt – Details of the data set, data quality, and reference.
5. tweets.csv – Data related to healthcare in the UK scraped from Twitter.

## PEP8 Guidelines

The PEP8 (Python Enterprise Proposal) is a document that provides various guidelines to write a readable code in Python. The code then becomes easy for anyone reading the code at a later stage. These guidelines were incorporated in the coding process.

## Import Libraries

To begin with, we import the main libraries required for loading, cleaning, and analyzing the data in Python. Those libraries are as follows:

- Pandas: Pandas stands for "Python Data Analysis Library". Pandas is a powerful and flexible open-source data analysis and manipulation tool.
- NumPy: adds support for large, multi-dimensional arrays and matrices, along with a large collection of high-level mathematical functions to operate on these arrays.
- Matplotlib: It is a plotting library for the Python programming language and its numerical mathematics extension NumPy.
- Seaborn: Seaborn is a Python data visualization library based on matplotlib.
- datetime : From datetime we import the date object we will require for manipulating the date objects in the dataset like appointment date.

## Loading the Data

Three main DataFrames are needed to store the data sets. DataFrame is a 2-dimensional labelled data structure with columns of potentially different types, like a spreadsheet.

We can view the DataFrame and see the columns and the rows in it as shown below.

```
.3]: # View the DataFrame.
     nc
```

.3]:

| | appointment_date | icb_ons_code | sub_icb_location_name | service_setting | context_type | national_category | count_of_appointments | appointment_month |
|---|---|---|---|---|---|---|---|---|
| 0 | 2021-08-02 | E54000050 | NHS North East and North Cumbria ICB - 00L | Primary Care Network | Care Related Encounter | Patient contact during Care Home Round | 3 | 2021-08 |
| 1 | 2021-08-02 | E54000050 | NHS North East and North Cumbria ICB - 00L | Other | Care Related Encounter | Planned Clinics | 7 | 2021-08 |
| 2 | 2021-08-02 | E54000050 | NHS North East and North Cumbria ICB - 00L | General Practice | Care Related Encounter | Home Visit | 79 | 2021-08 |
| 3 | 2021-08-02 | E54000050 | NHS North East and North Cumbria ICB - 00L | General Practice | Care Related Encounter | General Consultation Acute | 725 | 2021-08 |
| 4 | 2021-08-02 | E54000050 | NHS North East and North Cumbria ICB - 00L | General Practice | Care Related Encounter | Structured Medication Review | 2 | 2021-08 |
| ... | ... | ... | ... | ... | ... | ... | ... | ... |
| 817389 | 2022-06-30 | E54000054 | NHS West Yorkshire ICB - X2C4Y | Extended Access Provision | Care Related Encounter | Unplanned Clinical Activity | 12 | 2022-06 |
| 817390 | 2022-06-30 | E54000054 | NHS West Yorkshire ICB - X2C4Y | Extended Access Provision | Care Related Encounter | Planned Clinics | 4 | 2022-06 |
| 817391 | 2022-06-30 | E54000054 | NHS West Yorkshire ICB - X2C4Y | Extended Access Provision | Care Related Encounter | Planned Clinical Procedure | 92 | 2022-06 |
| 817392 | 2022-06-30 | E54000054 | NHS West Yorkshire ICB - X2C4Y | Extended Access Provision | Care Related Encounter | General Consultation Routine | 4 | 2022-06 |
| 817393 | 2022-06-30 | E54000054 | NHS West Yorkshire ICB - X2C4Y | Extended Access Provision | Care Related Encounter | General Consultation Acute | 19 | 2022-06 |

We then sense check the dataset for missing values and describe the data in terms of its columns and its shape and view the basic descriptive statistics to understand the data.

```
: # Determine whether there are missing values.
  nc.isnull().sum()
```

```
: appointment_date         0
  icb_ons_code             0
  sub_icb_location_name    0
  service_setting          0
  context_type             0
  national_category        0
  count_of_appointments    0
  appointment_month        0
  dtype: int64
```

```
: # Determine the metadata of the data set.
  print("The metadata of the national_categories nc dataframe:")
  print(nc.info())
```

```
The metadata of the national_categories nc dataframe:
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 817394 entries, 0 to 817393
Data columns (total 8 columns):
 #   Column                Non-Null Count   Dtype
---  ------                --------------   -----
 0   appointment_date      817394 non-null  datetime64[ns]
 1   icb_ons_code          817394 non-null  object
 2   sub_icb_location_name 817394 non-null  object
 3   service_setting       817394 non-null  object
 4   context_type          817394 non-null  object
 5   national_category     817394 non-null  object
 6   count_of_appointments 817394 non-null  int64
 7   appointment_month     817394 non-null  object
dtypes: datetime64[ns](1), int64(1), object(6)
memory usage: 49.9+ MB
```

## EXPLORING THE DATA

The DataFrames are now ready to be explored and analyzed. For example, we use the DataFrame to view the locations in the national categories dataset or the different context types and service setting types etc. along with counts of unique values present in these columns.

```
: # Determine the number of context types.
  print("The number of context types with their record counts are :\n",nc.context_type.value_counts())
```

```
The number of context types with their record counts are :
 Care Related Encounter    700481
Inconsistent Mapping       89494
Unmapped                   27419
Name: context_type, dtype: int64
```

```
: # Determine the number of national categories.
  print("The different of national categories with their record counts are :\n",nc.national_category.value_coun
```

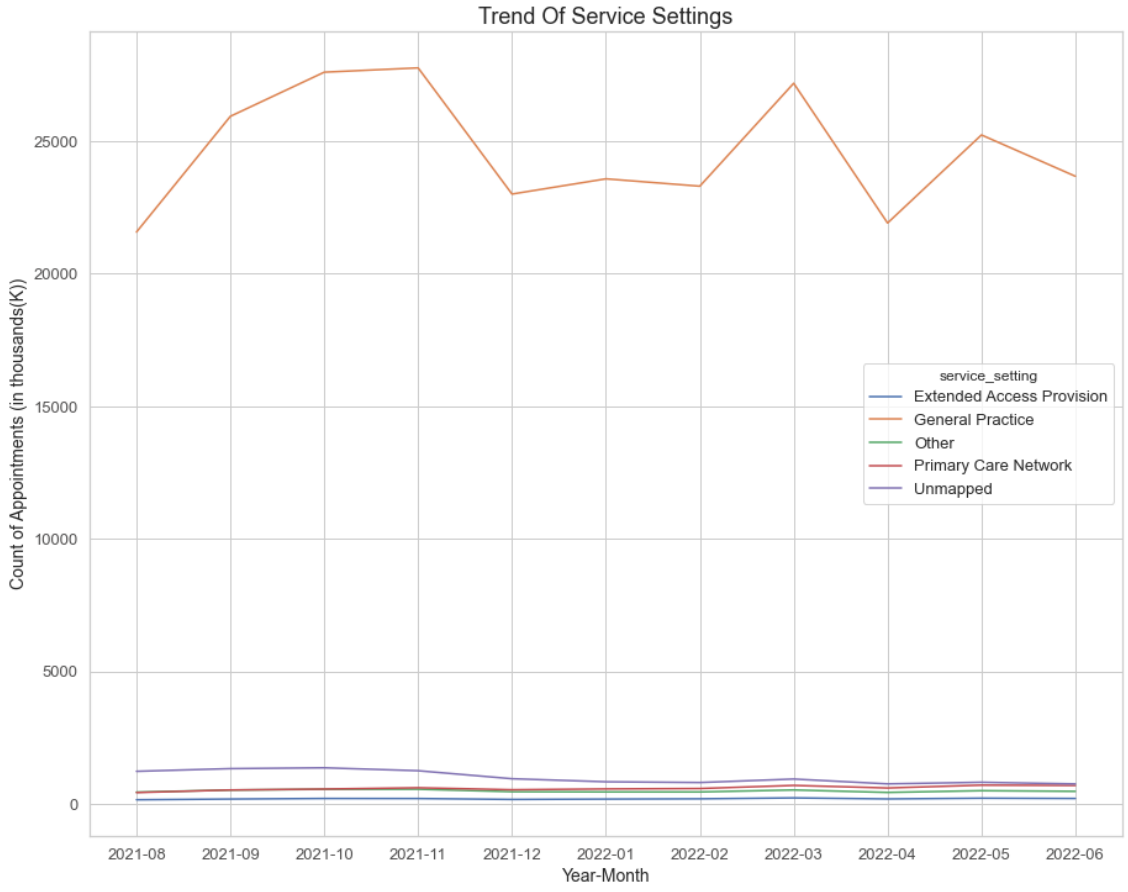
```
The different of national categories with their record counts are :
 Inconsistent Mapping                                                89494
General Consultation Routine                                         89329
General Consultation Acute                                           84874
Planned Clinics                                                      76429
Clinical Triage                                                      74539
Planned Clinical Procedure                                           59631
Structured Medication Review                                         44467
Service provided by organisation external to the practice            43095
Home Visit                                                           41850
Unplanned Clinical Activity                                          40415
Patient contact during Care Home Round                               28795
Unmapped                                                             27419
Care Home Visit                                                      26644
Social Prescribing Service                                           26492
Care Home Needs Assessment & Personalised Care and Support Planning  23505
Non-contractual chargeable work                                      20896
Walk-in                                                              14179
Group Consultation and Group Education                                5341
Name: national_category, dtype: int64
```

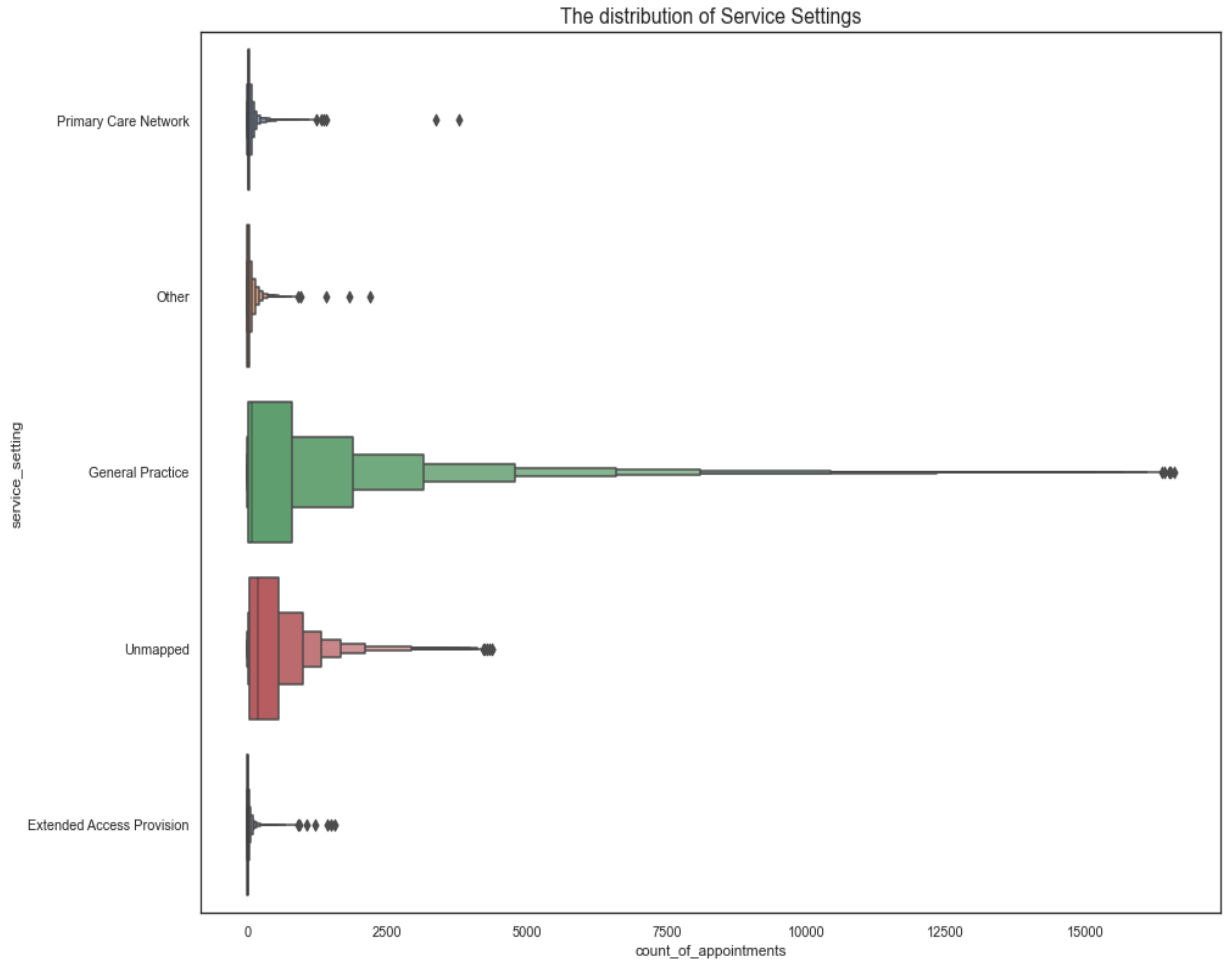
We perform further calculations and use the available aggregating methods to answer specific questions about the data. For example, we can infer using the DataFrames that the data collected for the actual duration dataset is between December,2021 to June 2022.The data collected for the national categories dataset is between August,2021 and June,2022. Choosing the 'NHS North West London ICB - W2U3Z' location (which is among the top five locations in terms of number of records present in national categories dataset) to study the appointments between the periods 1 January to 1 June 2022, we find that the ‘General Practice’ service setting was the most popular service setting. And overall, in the

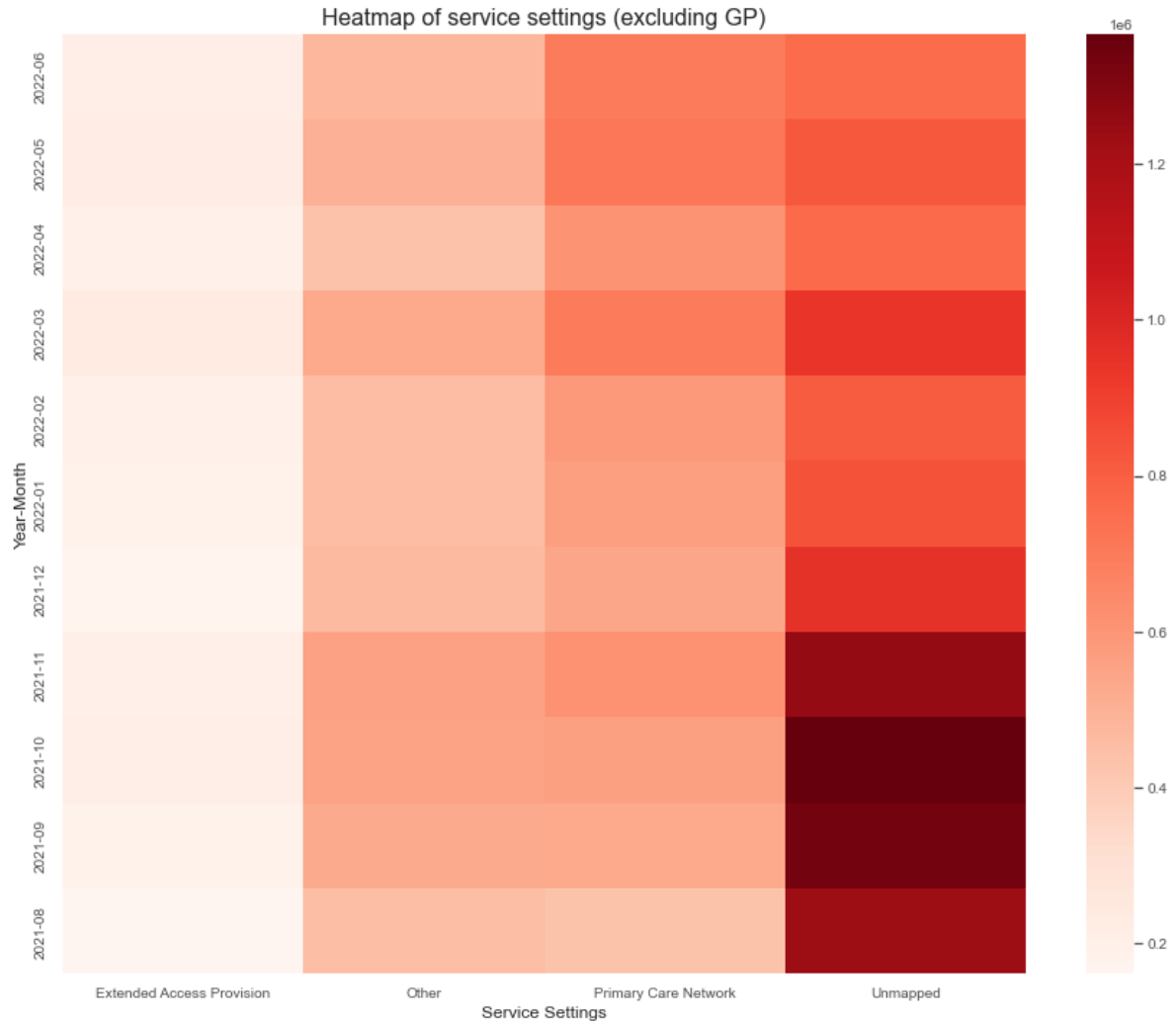
national categories dataset, November,2021 saw the highest number of appointments.

## VISUALISATION AND INSIGHTS

Next, we create different visualizations and identify possible monthly and seasonal trends in the data. For example, visualizing the service settings monthly trend, we can see that the 'General Practice' setting has a much higher count of appointments compared to other service settings. It therefore seems prudent to examine the General Practice setting separately. First, we visualize the distribution of service settings using a boxenplot in seaborn which is like a boxplot.

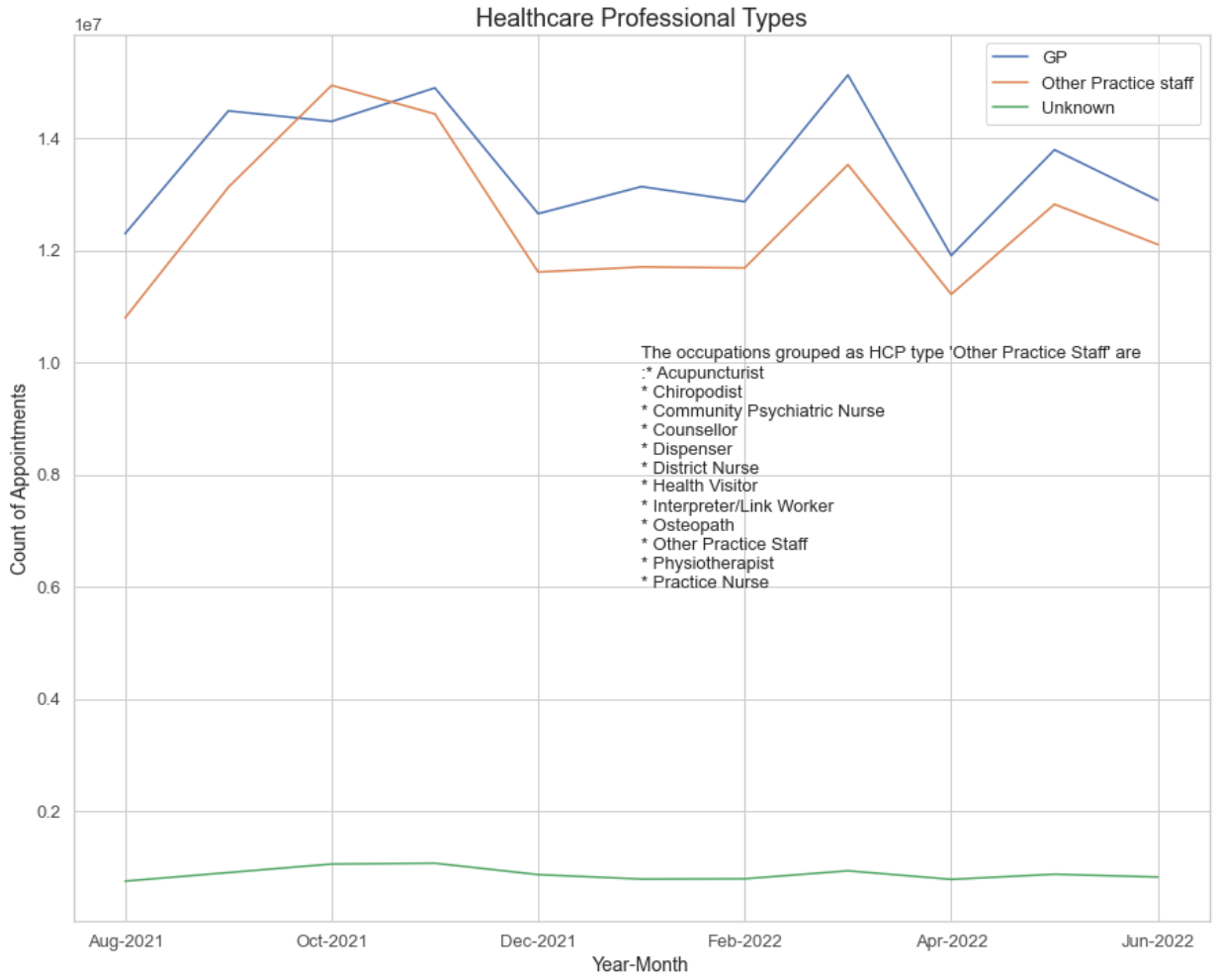



The distribution of Service Settings
Primary Care Network
Other
General Practice
Unmapped
Extended Access Provision
service_setting
0
2500
5000
7500
10000
12500
15000
count_of_appointments




Looking at the heatmap of other service settings excluding GP we can see that the Unmapped values are higher in October 2021 and around that period and it goes down in 2022. Below we see the trend of GP service setting separately.

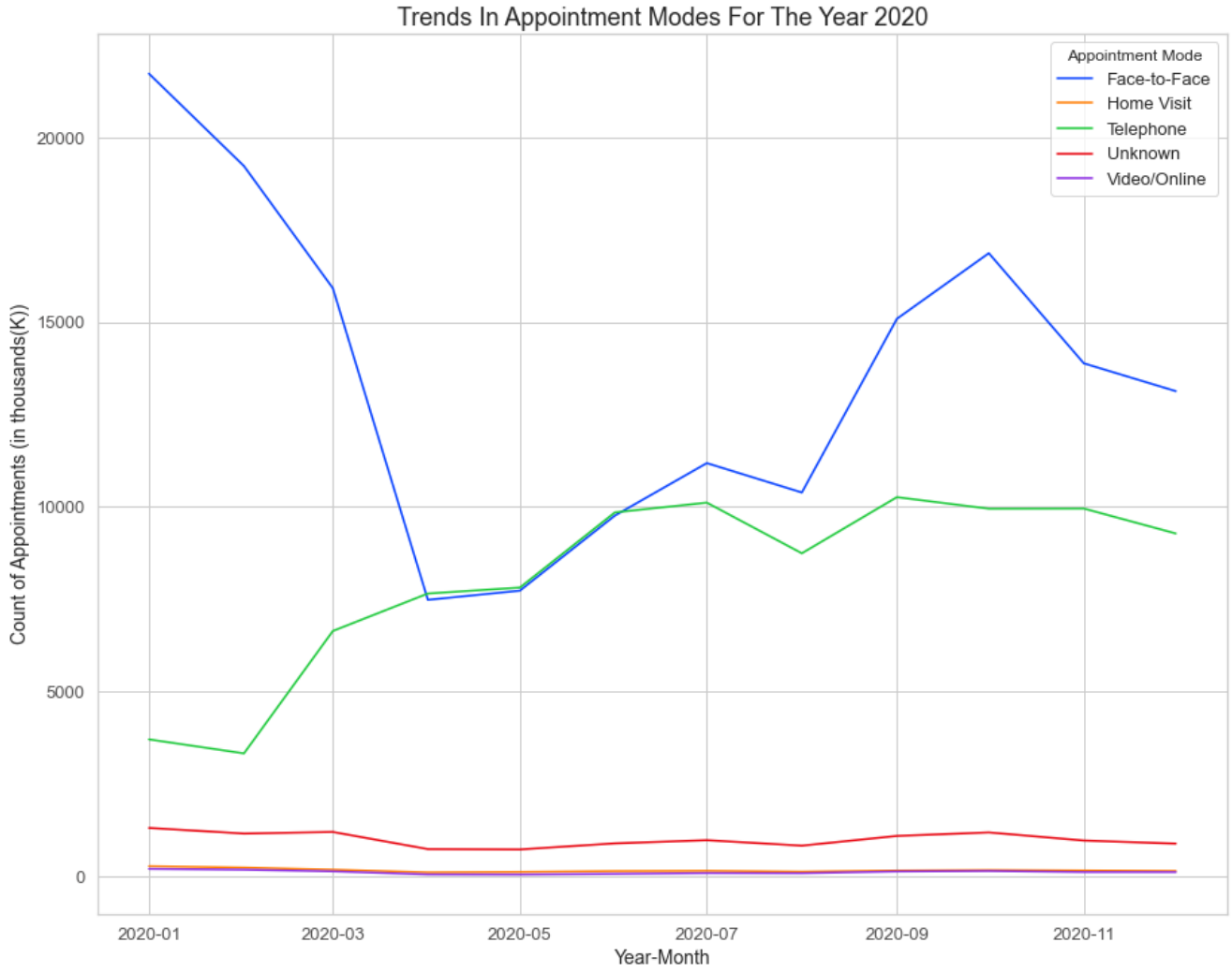

Similarly, we can visualize and explore each of the columns to make similar inferences about the trends present in the data. For example, the monthly trend in healthcare professional types -:

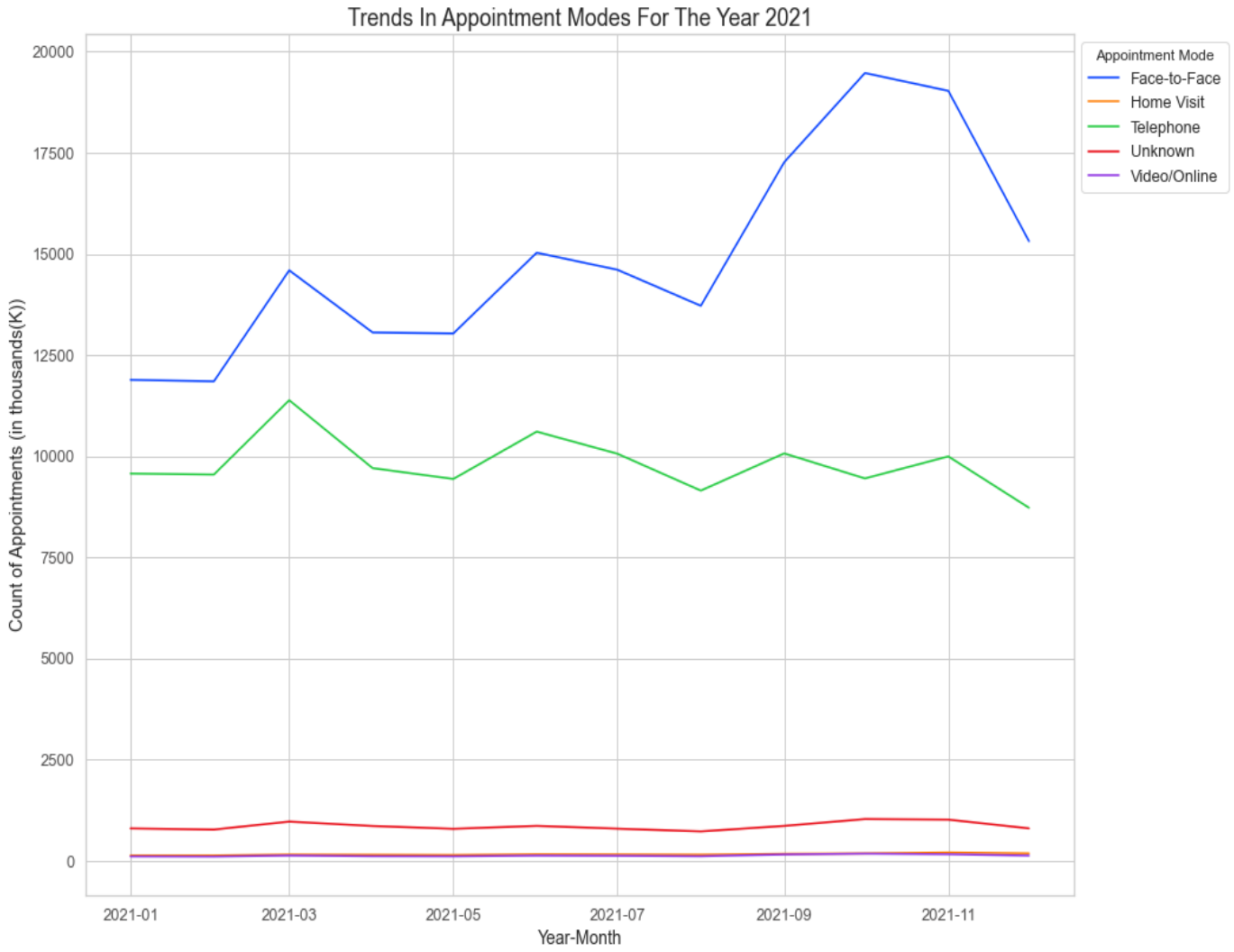


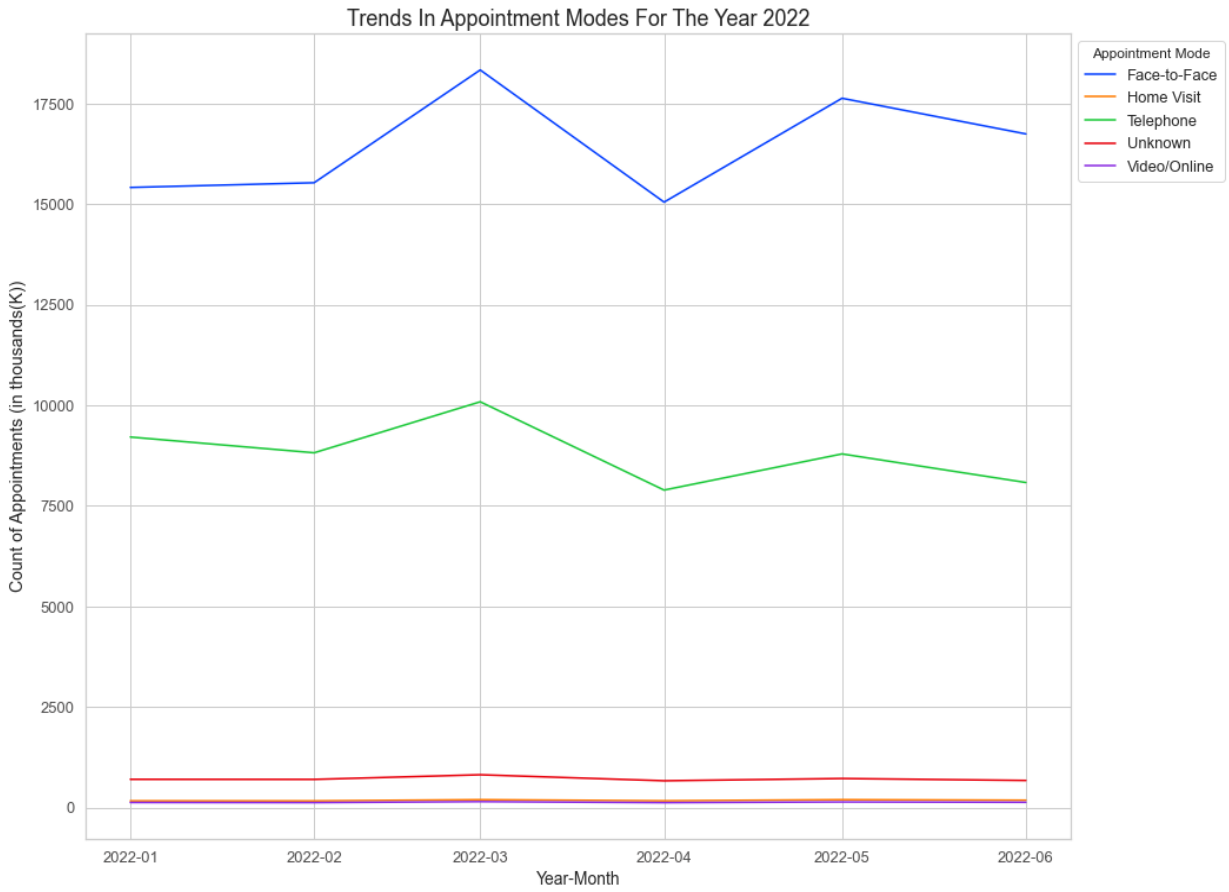

## PATTERNS AND PREDICTIONS

Looking at the trends in appointment modes for the years 2020, 2021,2022 based on available data we can see the dip in face-to-face appointments for the Covid lockdown period and then revert to its usual

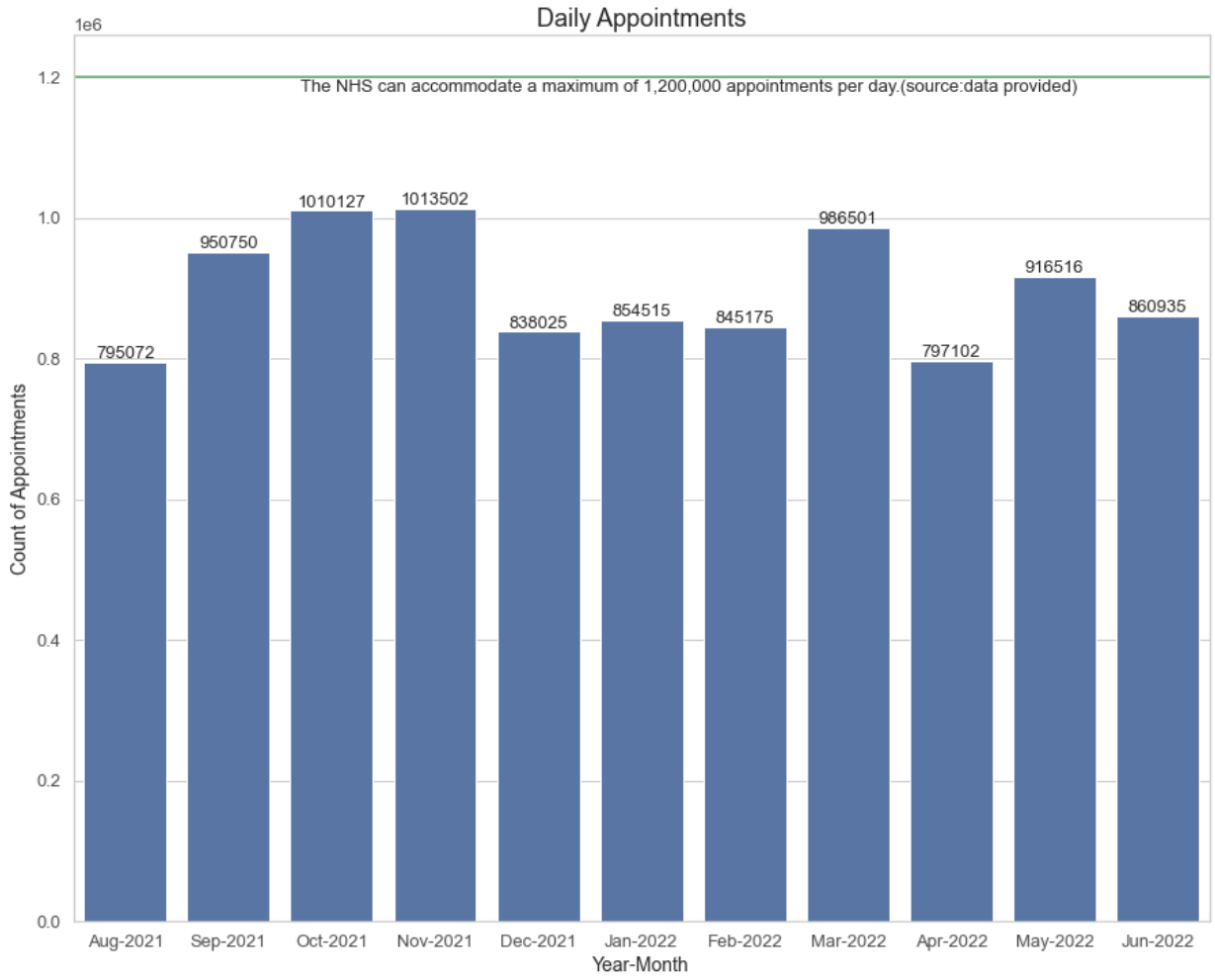
trend later.


Trends In Appointment Modes For The Year 2020
Appointment Mode
Face-to-Face
Home Visit
Telephone
Unknown
Video/Online
Count of Appointments (in thousands(K))
20000
15000
10000
5000
0
2020-01
2020-03
2020-05
2020-07
2020-09
2020-11
Year-Month

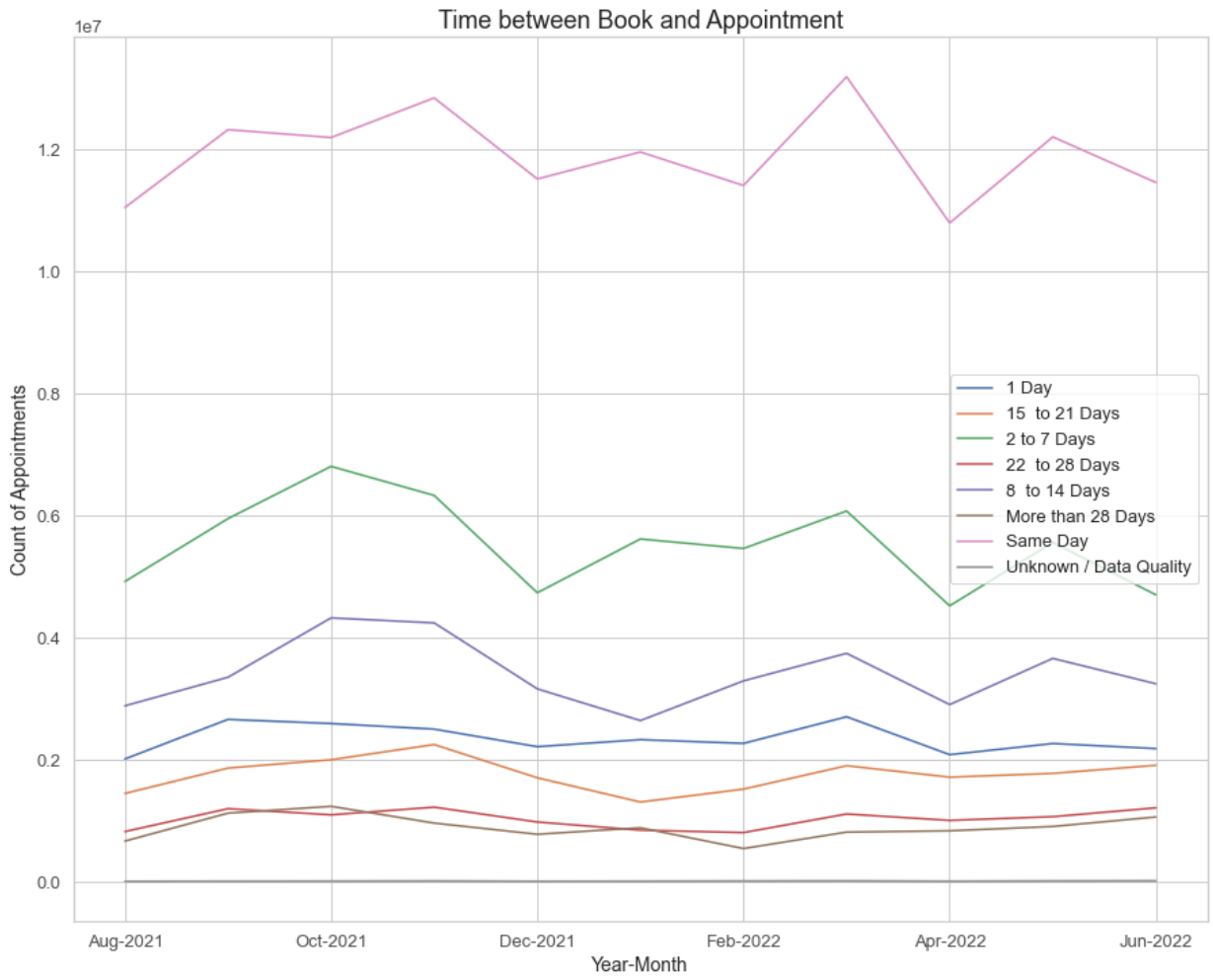

Trends In Appointment Modes For The Year 2021
Count of Appointments (in thousands(K))
20000
17500
15000
12500
10000
7500
5000
2500
0
2021-01
2021-03
2021-05
2021-07
2021-09
2021-11
Year-Month
Appointment Mode
Face-to-Face
Home Visit
Telephone
Unknown
Video/Online




Also based on the information provided the NHS can accommodate a maximum of 1,200,000 appointments per day. We aggregate the monthly data and divide by 30 to get an approximate daily estimate of appointments in the national categories dataset.

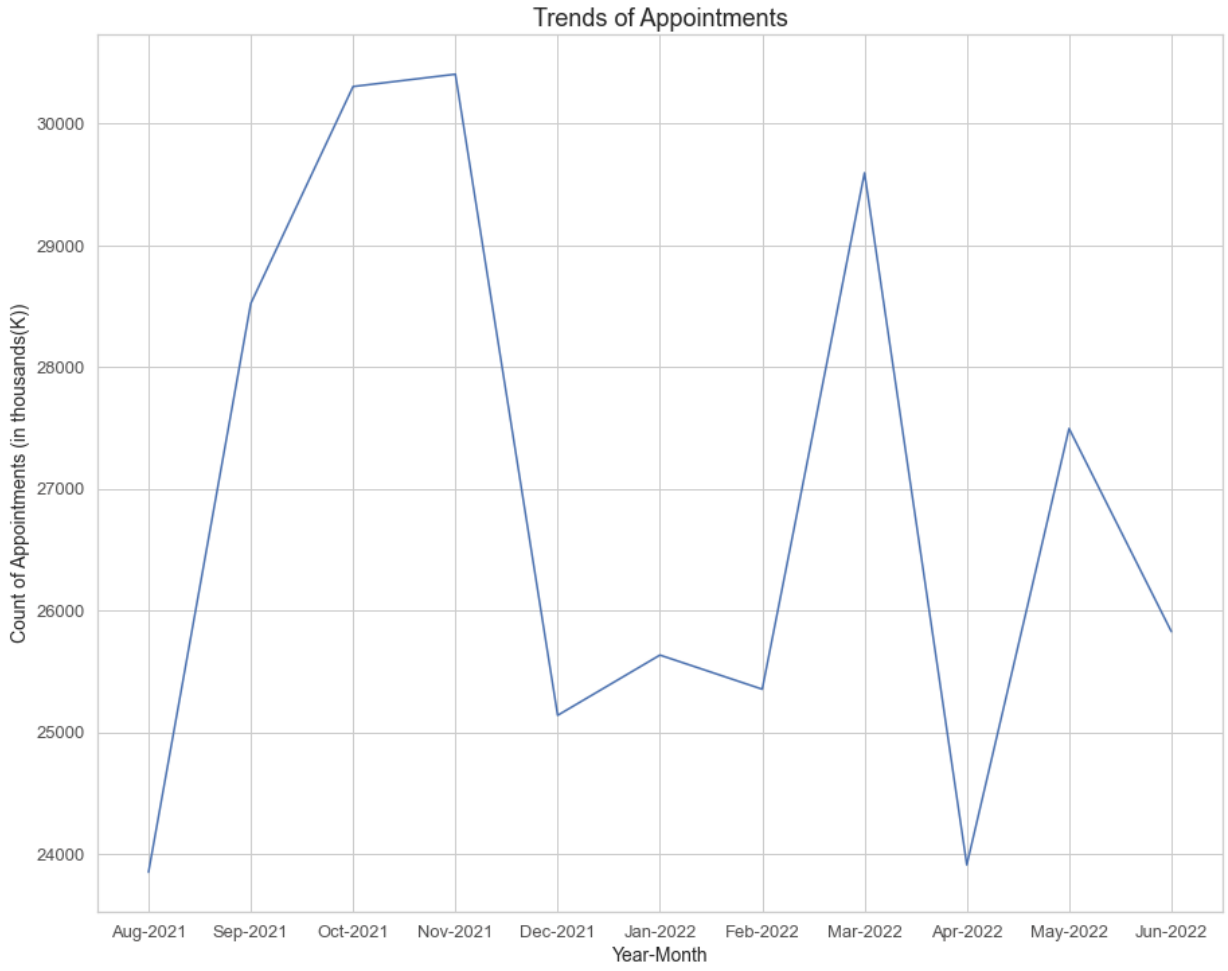



Furthermore, looking at the data we see that most appointments are booked for the same day.

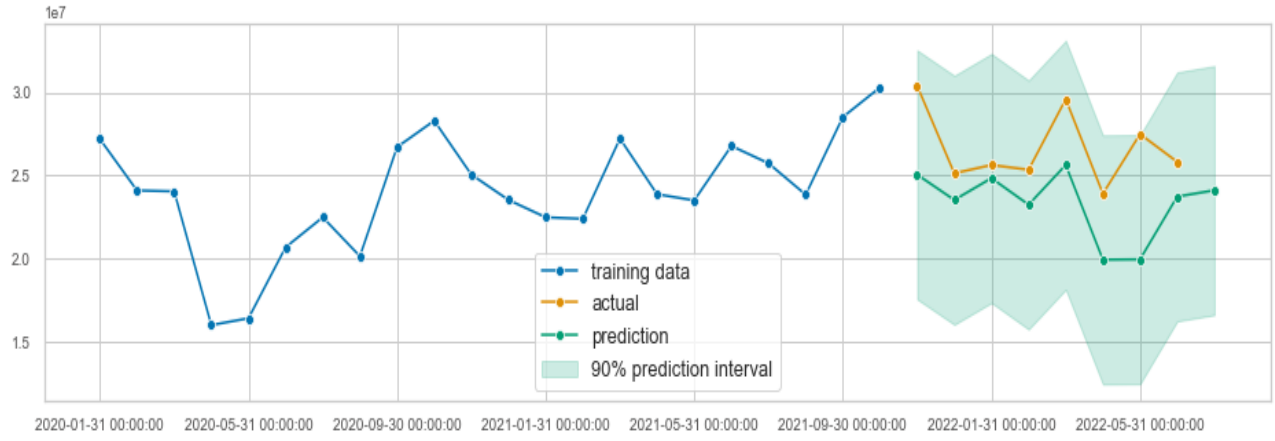



### *Prediction*

We can visualize the trend of appointments using the appointments regional dataset and then use a package **sktime** designed to analyze time series trends and make a forecast to predict future appointments.




The Naive Forecaster was chosen here because it provided the best results. The naive forecaster assumes the past trends will continue and forecasts on that basis. The mean absolute percentage error value i.e., Mape value (0.125) and the predicted intervals show that the forecast is acceptable but could be more accurate as more data is accumulated and less affected by the steep decline corresponding to the 2020 lockdown period.



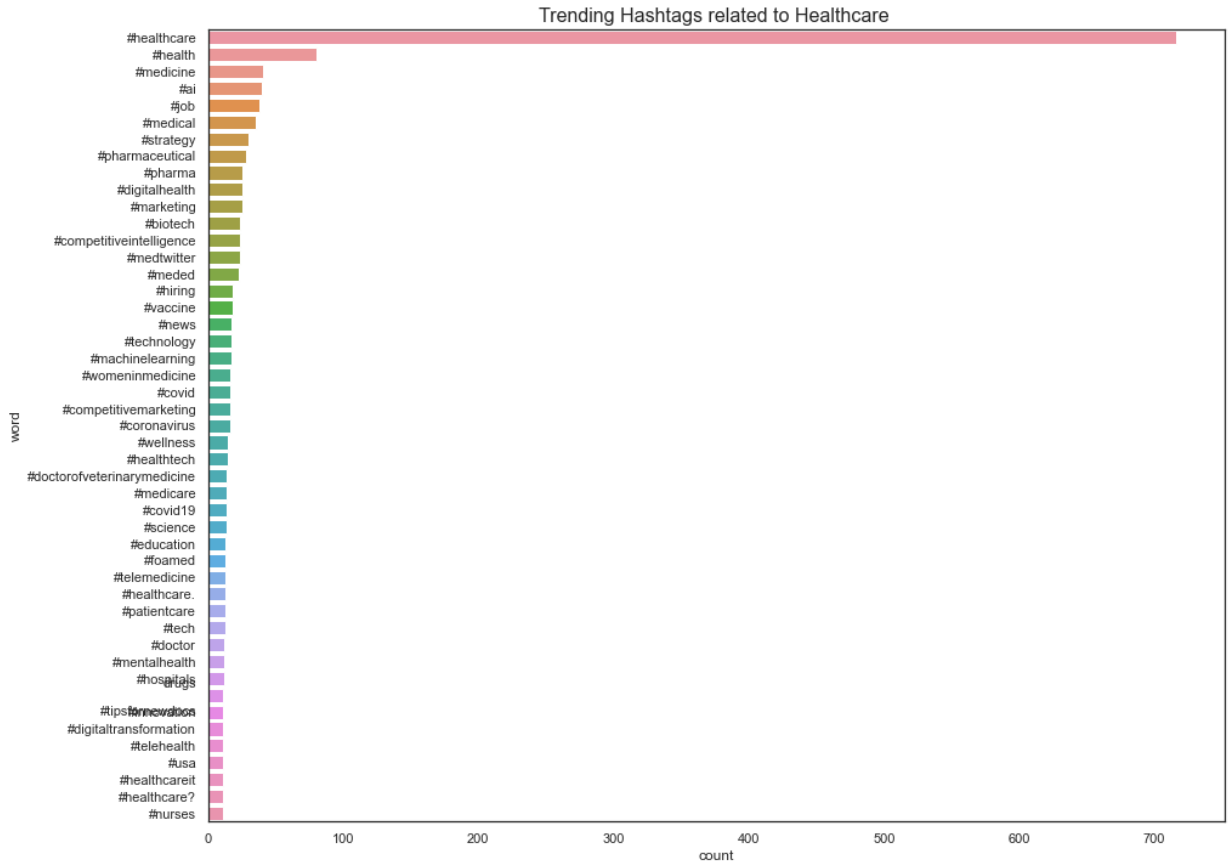

## ANALYZING THE TWITTER DATA

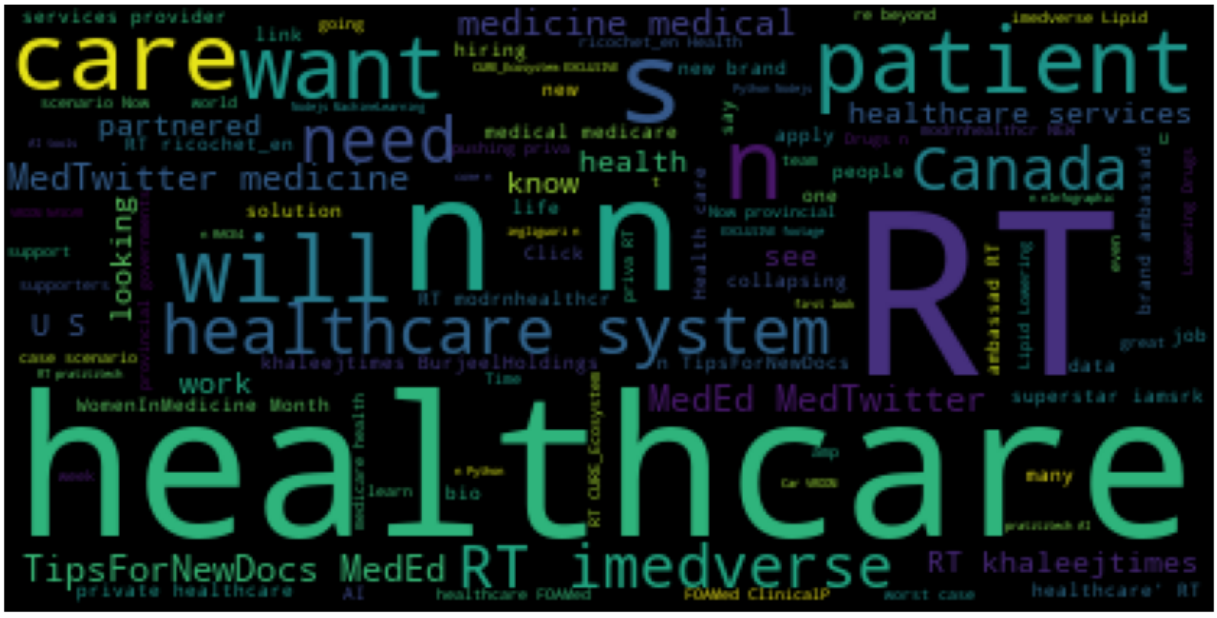
We identified the trending hashtags and related to healthcare in UK based on the set of tweets received.

Here is a sample of the top favorited tweets.

| | | | | |
|---|---|---|---|---|
| **442** | 1567634936341069826 | How health insurance works 😂 \n\n#comedy #adulting #healthcare https://t.co/ciksdeoAkb | {'hashtags': [{'text': 'comedy', 'indices': [31, 38]}, {'text': 'adulting', 'indices': [39, 48]}, {'text': 'healthcare', 'indices': [49, 60]}], 'symbols': [], 'user_mentions': [], 'urls': [], 'med... | #comedy, #adulting, #healthcare |
| **84** | 1567579049043832832 | Our nat'l choices re: #healthcare systems aren't the continuum of public or private, but how much we want of:\n\n- fragmented or seamless\n- does simplicity or complexity well\n- prioritizes savin... | {'hashtags': [{'text': 'healthcare', 'indices': [22, 33]}, {'text': 'cdnpoli', 'indices': [270, 278]}], 'symbols': [], 'user_mentions': [], 'urls': []} | #healthcare, #cdnpoli |
| **1122** | 1567586306607423488 | Heart Failure, Myocardial Infarction & immediate Treatment\n\n#TipsForNewDocs #MedEd #MedTwitter #medicine #medical #medicare #health #healthcare #FOAMed #ClinicalPearl #clinicaltips #MedStude... | {'hashtags': [{'text': 'TipsForNewDocs', 'indices': [64, 79]}, {'text': 'MedEd', 'indices': [80, 86]}, {'text': 'MedTwitter', 'indices': [87, 98]}, {'text': 'medicine', 'indices': [99, 108]}, {'te... | #TipsForNewDocs, #MedEd, #MedTwitter, #medicine, #medical, #medicare, #health, #healthcare, #FOAMed, #ClinicalPearl, #clinicaltips, #MedStudents, #medstudenttwitter, #lipic |
| **119** | 1567577266162475011 | More data that our 13+ 🇨🇦 #healthcare systems fall short of providing adequate access to care, even for those w/ a family physician.\n\nInstead of siloed solutions and further system fragmentation... | {'hashtags': [{'text': 'healthcare', 'indices': [26, 37]}], 'symbols': [], 'user_mentions': [], 'urls': [{'url': 'https://t.co/eZkz5brcYT', 'expanded_url': 'https://www.ctvnews.ca/health/canadians... | #healthcare |
| **758** | 1567611240024875008 | Looking forward to speaking at #ConV2X on Sep 15! Ping me for a speaker discount if interested! Register at https://t.co/v20ebbXmdO \n\n.@BHTYjournal @hedera @acoerco\n#blockchain #DLT #healthcare... | {'hashtags': [{'text': 'ConV2X', 'indices': [31, 38]}, {'text': 'blockchain', 'indices': [165, 176]}, {'text': 'DLT', 'indices': [177, 181]}, {'text': 'healthcare', 'indices': [182, 193]}, {'text'... | #ConV2X, #blockchain, #DLT, #healthcare, #innovation |
| **1098** | 1567587492949286912 | @CapricornFMNews We have waiting to hear this kind of news now Sa is getting things correct there is absolutely nothing mahala #HealthCare services in SA must be paid by foreign nationals and loca... | {'hashtags': [{'text': 'HealthCare', 'indices': [127, 138]}], 'symbols': [], 'user_mentions': [{'screen_name': 'CapricornFMNews', 'name': 'CapricornFM News', 'id': 1481829344, 'id_str': '148182934... | #HealthCare |
| **342** | 1567643206480699392 | September is #WomenInMedicine Month! Thrilled to join @JulieSilverMD #SheLeadsHealthcare @ELAMProgram & @AMWADoctors to #InvestInHer!\n\nI #InvestInHer to diversify the #healthcare #leadersh... | {'hashtags': [{'text': 'WomenInMedicine', 'indices': [13, 29]}, {'text': 'SheLeadsHealthcare', 'indices': [70, 89]}, {'text': 'InvestInHer', 'indices': [126, 138]}, {'text': 'InvestInHer', 'indice... | #WomenInMedicine, #SheLeadsHealthcare, #InvestInHer, #InvestInHer, #healthcare, #leadership, #racialequity, #genderequity, #MedTwitter, #HeForShe |

The trending hashtags visualized as a graph and in a wordcloud using the wordcloud package .



## SUMMARY AND RECOMMENDATIONS

In summary, we can see that General Practice service setting has the highest number of records with most of the appointments being booked the same day and face-to-face mode is preferred.

The monthly trends in healthcare professional types combined with the forecast of future appointments can then be used to estimate staffing requirements. Steps should be taken to minimize inconsistent mapping and unmapped values. The daily appointment graph suggests more appointments can be accommodated daily and a more in-depth study could be done to analyze the issues surrounding effective utilization.