Assignment: Diagnostic Analysis using Python

By Timothy Ayling

Background/context of the business:

Contracted by the NHS to explore the reduction of missed appointments through a data informed approach.

NHS questions:

- Adequate staff and capacity in the networks?
- What was the actual utilisation of resources?

Investigative analysis:

1. What is the number of locations, service settings, context types, national categories, and appointment statuses in the data sets?
2. What is the date range of the provided data sets, and which service settings reported the most appointments for a specific period?
3. What is the number of appointments and records per month?
4. What monthly and seasonal trends are evident, based on the number of appointments for service settings, context types, and national categories?
5. What are the top trending hashtags (#) on Twitter related to healthcare in the UK?
6. Were there adequate staff and capacity in the networks?
7. What was the actual utilisation of resources?
8. What possible recommendations does the data provide for the NHS?

Analytical Approach:

I opened GitHub and opened a new repository selecting the public setting. I saved my Jupiter files (py and ipynb) in the attachments and continued to do so whilst working on my assignment. I eventually had three files saved there namely my report pdf file, my Jupyter Notebook ipynb file and my presentation recording in mp4. The 'Public option' in GitHub allows for a shared function available for team members and others to work from and use the data. I also gave further description to the file for easier selection and search parameters.

I imported the csv and xlsx extension files into Jupyter through the upload function. I then imported all the relevant libraries right at the beginning so my coding would not be limited in any way. I gave headers to each parameter coding line to give insight to the reader my action.

I then wrote code in the program to read it in a data frame. I then manipulated the data to give me further information i.e. shape dtypes, columns, head, tail, missing values. This is done for me to check the column names, the layout of the file, the number of rows and columns. The head(5) is used especially instead of looking at the whole spreadsheet which gives a sneak peek of the first 5 lines.

I determined the number of locations, service settings, context types, national categories and appointment statuses in the data sets.

I determined the number of appointments and records per month. I then determined the monthly and seasonal trends based on the figures I achieved from the above outcomes.

Value counts function was used extensively which returns a value to the number of occurrences in the data set. This assisted me in answering the questions. The date format had to be changed for time based calculations.

Datetime(), groupby(), loc() were functions I used to manipulate the data. Groupby() takes data from a larger set and deals with the selected data from your parameters to bring about order and understanding resulting in answering the questions.

Twitter: Max and min value and I sorted the data in ascending order. Data scraping techniques were implemented to extrapolate the data.

Visualisations and insights: 350

Through the analysis a large difference was recorded between the East and South East of the country and to the remainder. The majority of the trusts are recorded in these two areas whilst London only has one trust. A large percentage was shown to patients making the booking but not attending.

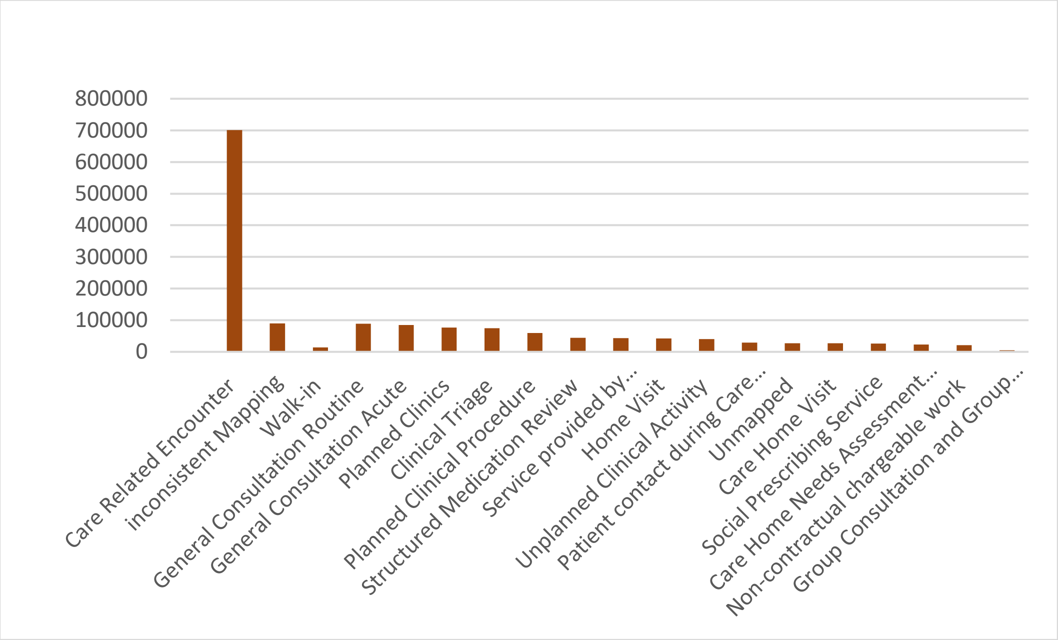
Some concrete conclusions were difficult to answer because of the unknown errors in the records. The recording element in the various medical offices need to come together and possibly use software that integrates with everyone or the same software or a shared database.

I have included below a visualisation “Appendix A” of the unknown value counts. Value_counts()

ar['appointment_status'].value_counts()

Appendix A

Care Related Encounter	700481
inconsistent Mapping	89494
Walk-in	14179
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
Group Consultation and Group Education	5341

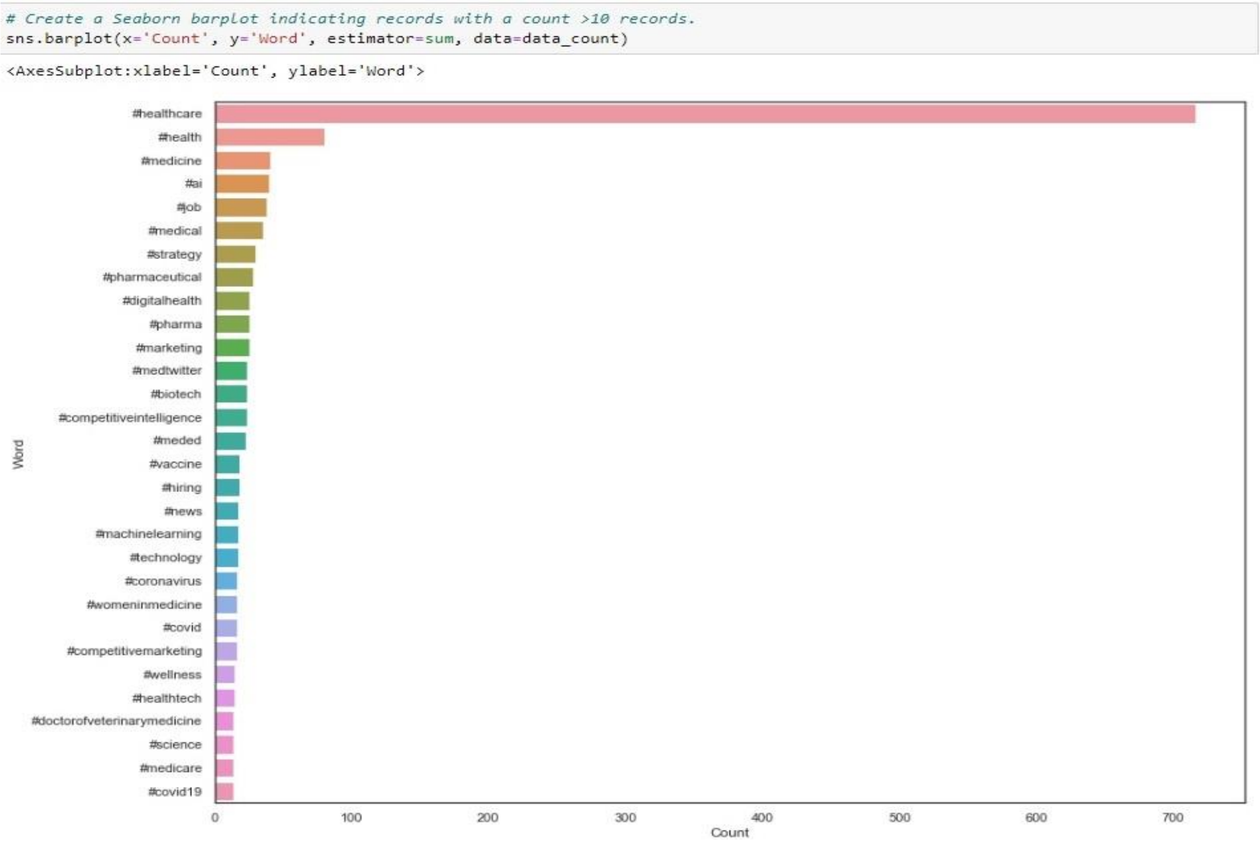


I set the display size and then used the groupby() function to select the columns that were relevant to the analysis. Codes were used form the library Seaborn with “sns”(an alias for Seaborn) to create line-plot and bar plots. The visualisations indicate the number of people attending GP surgeries are the highest, and it has been noted that GP surgeries can offer more than just consultation and can act as a community healthcare point. Visualisations were created for each question, splitting them further for analysis to make it clearer for the stakeholders which makes everything simpler.

As social media becomes the communication norm and more and more information about products and services can be had, I decided to scrap Twitter for any hashtags that are trending about NHS and in general about any particular service. The data scrap revealed that #healthcare is the most prominent and that shows many a views were put forward under this hashtag.

Please see the visual below noted Appendix B. I scanned for other hashtags by removing the popular ones and revealed the other hashtags that are trending but not as much as #healthcare. I then used the “remove outlier method” which eliminated some of the hashtags and displayed the rest.

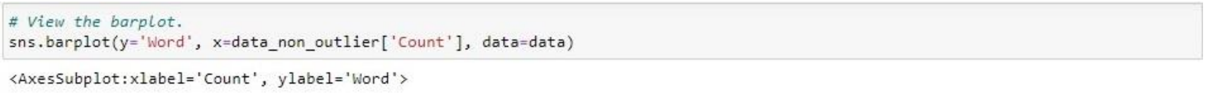
Appendix B



```
# Create a Seaborn barplot indicating records with a count >10 records.
sns.barplot(x='Count', y='Word', estimator=sum, data=data_count)
```

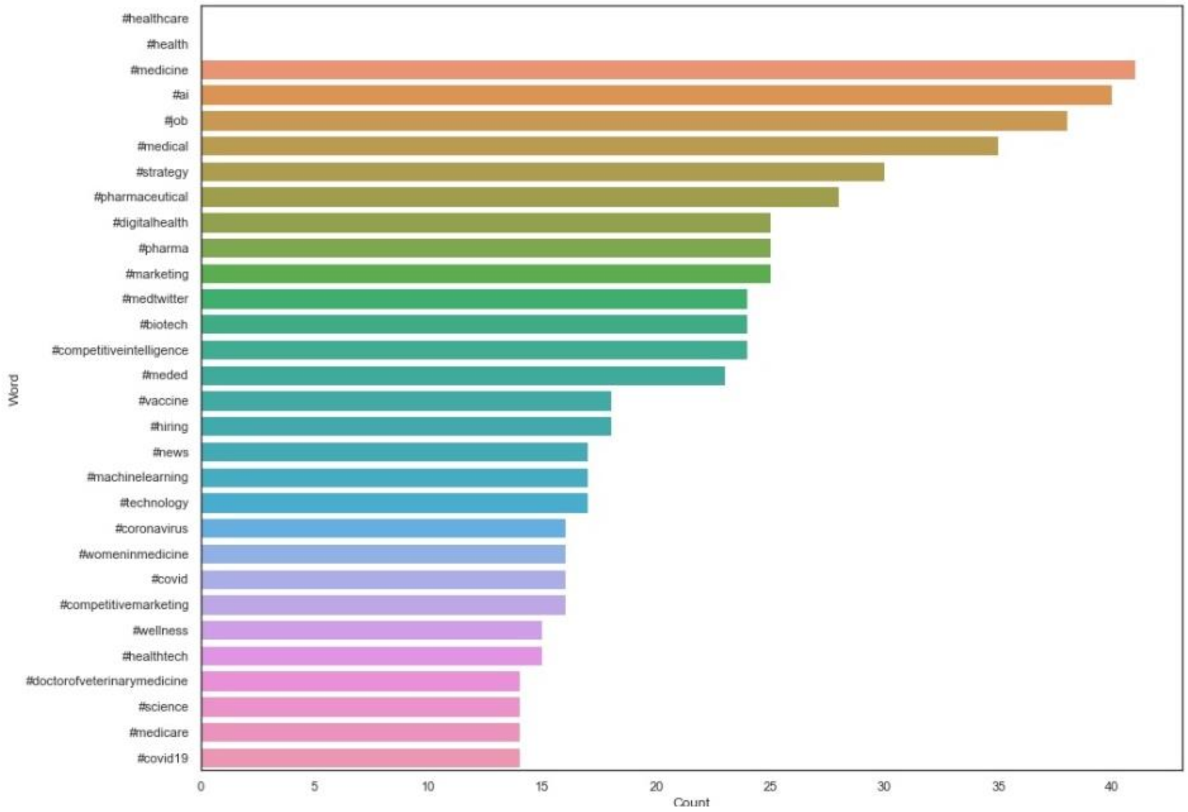
```
<AxesSubplot:xlabel='Count', ylabel='Word'>
```

The above visual is the #healthcare chart.



```
# View the barplot.
sns.barplot(y='Word', x=data_non_outlier['Count'], data=data)
```

```
<AxesSubplot:xlabel='Count', ylabel='Word'>
```



The visual above is without #healthcare to give a clearer understanding of other trending hashtags.

Patterns /predictions and Summary:

As indicated clearly from the visuals the highest number of patients with regarding appointments was on the day and seeing the GP face to face and the A and E staff. This would interpret as increased pressure for the staff, longer queues for the patients for many hours.

Also looking at another visual, the NHS peak times in the year came in the season of Winter and Summer.

The data sets provided also show that some trusts are overwhelmed with patients whilst others are underwhelmed. This comes into one of the questions where it speaks about the utilisation of resources.

The management of services offered at the hospitals need to be looked at to share the load. Possibly to upgrade the facilities underutilised. Also if anything can be done to reduce the queues in A and E in services offered elsewhere or greater numbers of staff provided.

More medical staff is required for the NHS to improve health, less anxiety, wellbeing and less suffering.

PS: The data that was given had a number of anomalies so to the best of extrapolation and isolation of data these were the results.