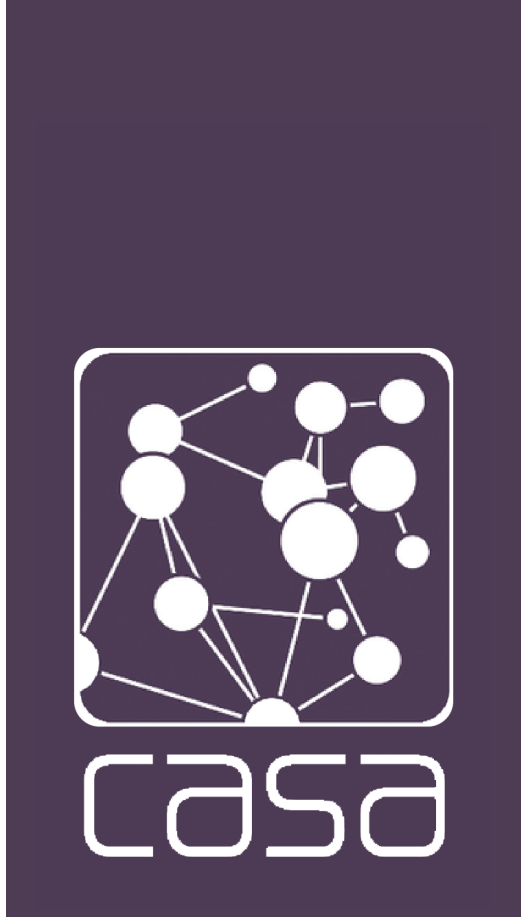


Exploring disease prevalence in England using clustering

Group 5

Antonios Fiala
Sara Moatti
Yating Fan

0

Content

1

Overview

1 Overview

- Explore the NHS data sets, here.
- Different geographical levels in the UK, focuses on several diseases
- Test the variation by doing a cluster analysis on the STP level
- The aim: identifying disease clusters that can guide future policies and decision making in public health area.
- We will now talk you through: data presentation, cluster methods, conclusion

2

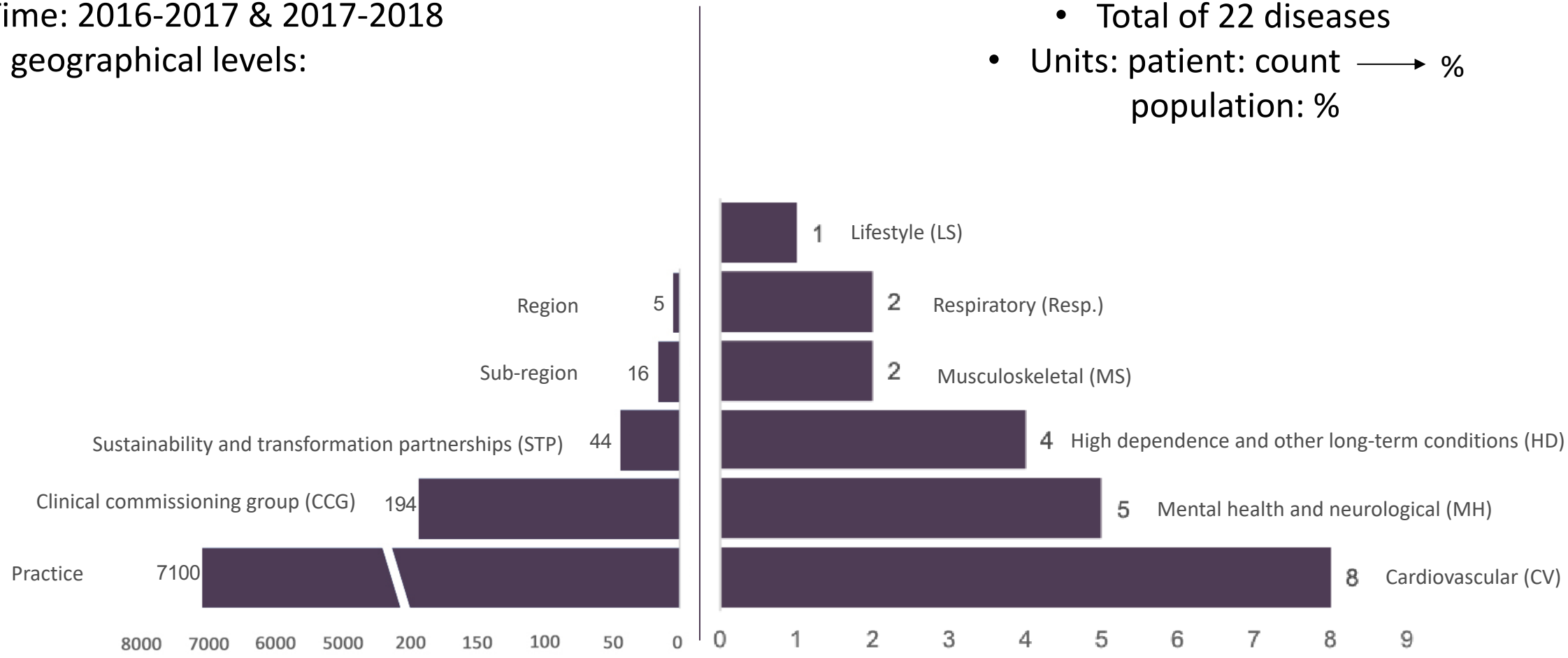
Data Presentation

- High-level data summary
- Variable selection
- Summary statistics

2 High level data summary

- Data source: NHS Digital website
- Time: 2016-2017 & 2017-2018
- 5 geographical levels:

- 6 categories of diseases
  - Total of 22 diseases
- Units: patient: count ⟶ %
  population: %

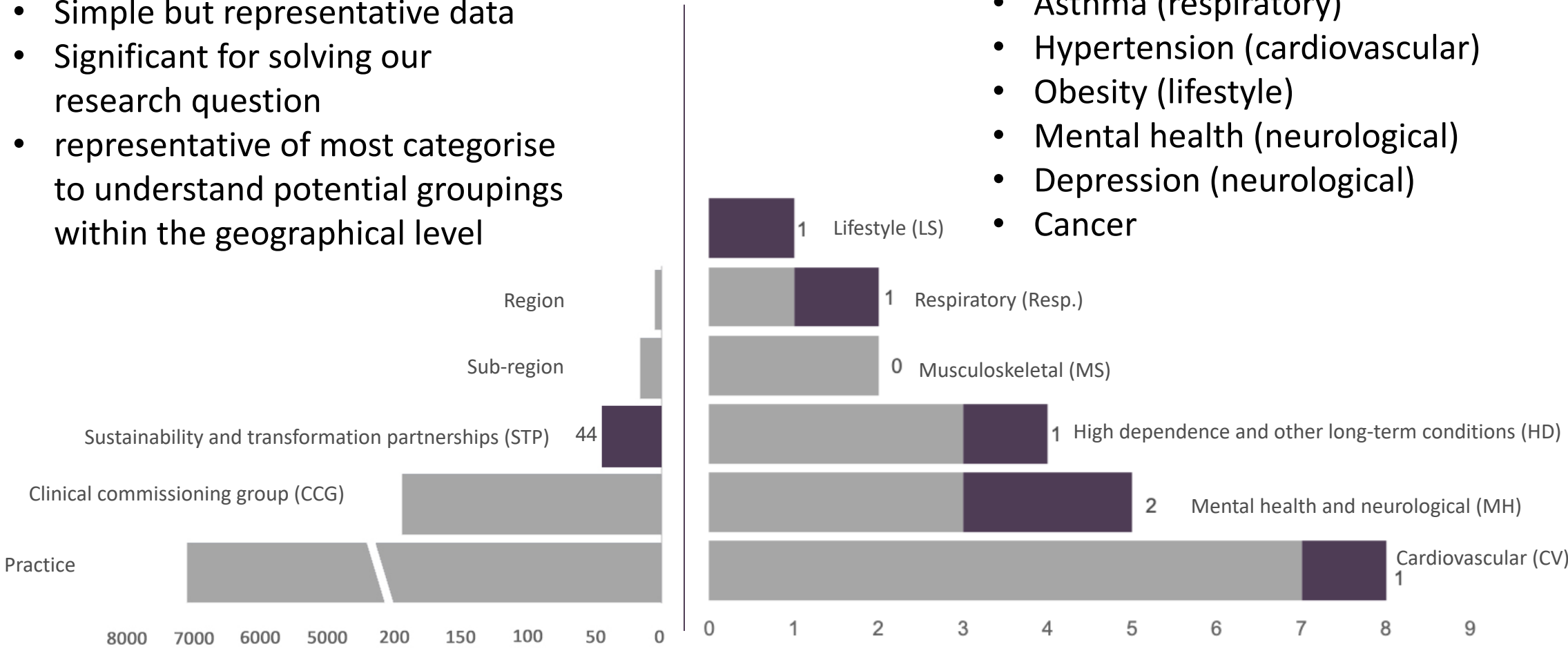


2 What we selected and why?

Geographical level : STP

- Simple but representative data
- Significant for solving our research question
- representative of most categorise to understand potential groupings within the geographical level

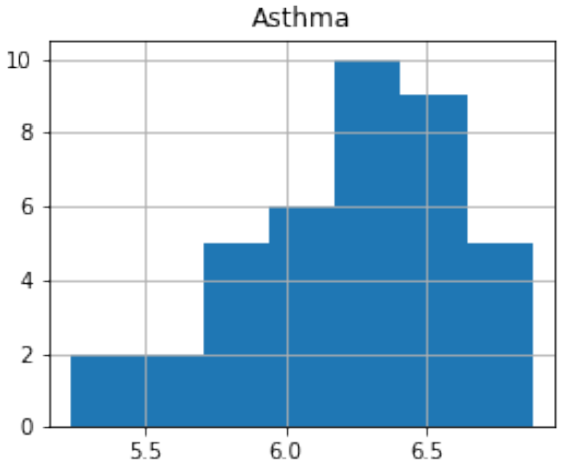
- 6 variables: Conditions
  - Asthma (respiratory)
  - Hypertension (cardiovascular)
  - Obesity (lifestyle)
  - Mental health (neurological)
  - Depression (neurological)
  - Cancer

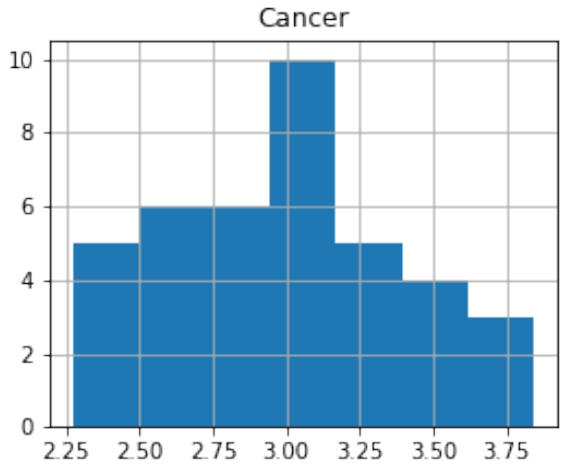


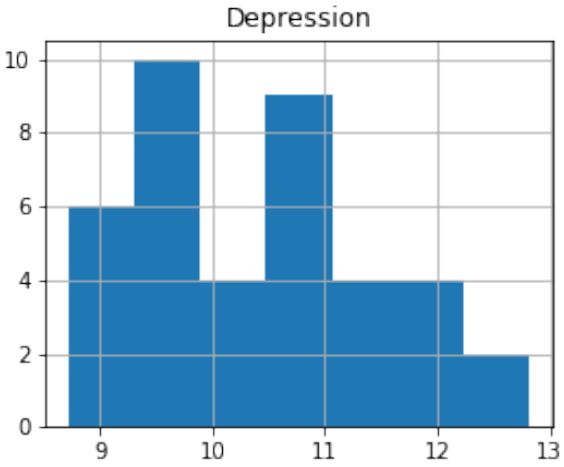
2 Summary statistics (a)

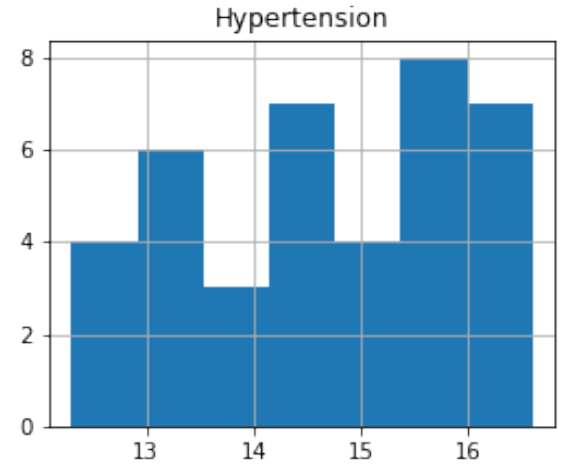
	Hypertension	Obesity	Cancer	Depression	Mental health	Asthma
count	44.000000	44.000000	44.000000	44.000000	44.000000	44.000000
mean	14.245036	9.794524	2.852103	10.007908	0.912965	6.040094
std	1.650992	1.722609	0.521830	1.444589	0.123327	0.668985
min	10.560932	5.820715	1.525505	6.384569	0.722217	4.389151
25%	13.162746	8.540744	2.569907	9.333540	0.817965	5.807491
50%	14.423130	9.717108	2.951305	10.027425	0.867090	6.265985
75%	15.520476	11.150265	3.200263	10.915953	0.996901	6.502184
max	16.600187	13.272109	3.840752	12.820154	1.267849	6.878169

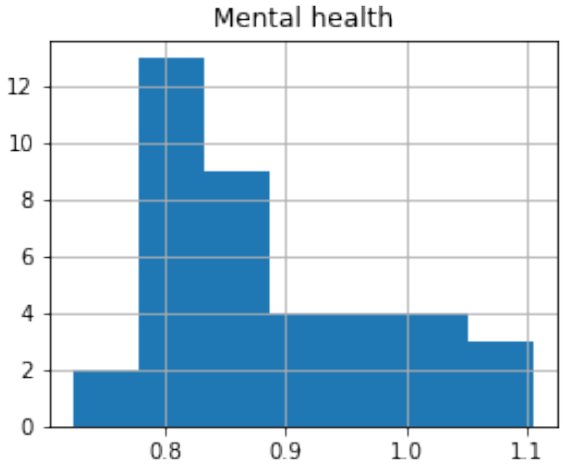
2 Data Distribution

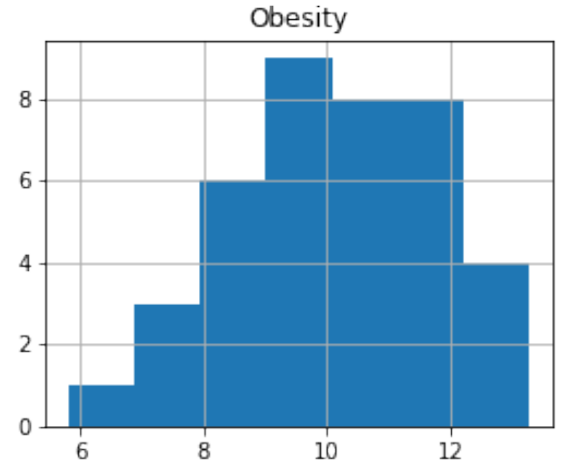






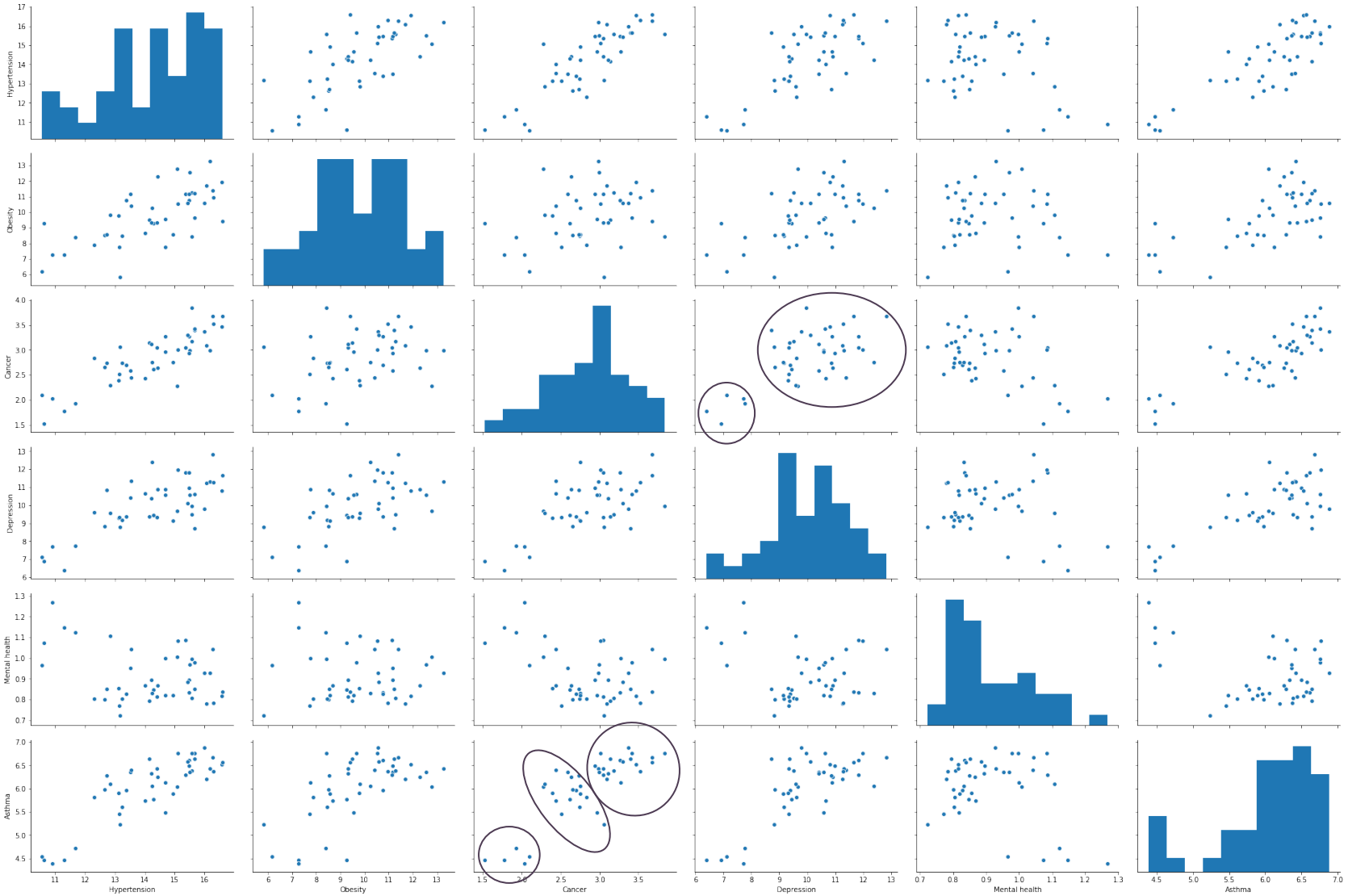






2

Visualisations



2 Summary statistics (b) - standardised

$$z = \frac{x - \mu}{\sigma}$$

	Hypertension	Obesity	Cancer	Depression	Mental health	Asthma
count	44	44	44	44	44	44
mean	3.22974e-16	-5.24833e-16	1.14508e-15	2.83864e-16	-7.36784e-16	1.34615e-15
std	1.01156	1.01156	1.01156	1.01156	1.01156	1.01156
min	-2.25725	-2.33352	-2.57159	-2.53721	-1.56457	-2.49637
25%	-0.663118	-0.736253	-0.547034	-0.47222	-0.779215	-0.351715
50%	0.109118	-0.0454608	0.192302	0.0136673	-0.376274	0.341565
75%	0.781461	0.796127	0.674904	0.635851	0.688472	0.698718
max	1.443	2.04213	1.91648	1.96925	2.91087	1.26724

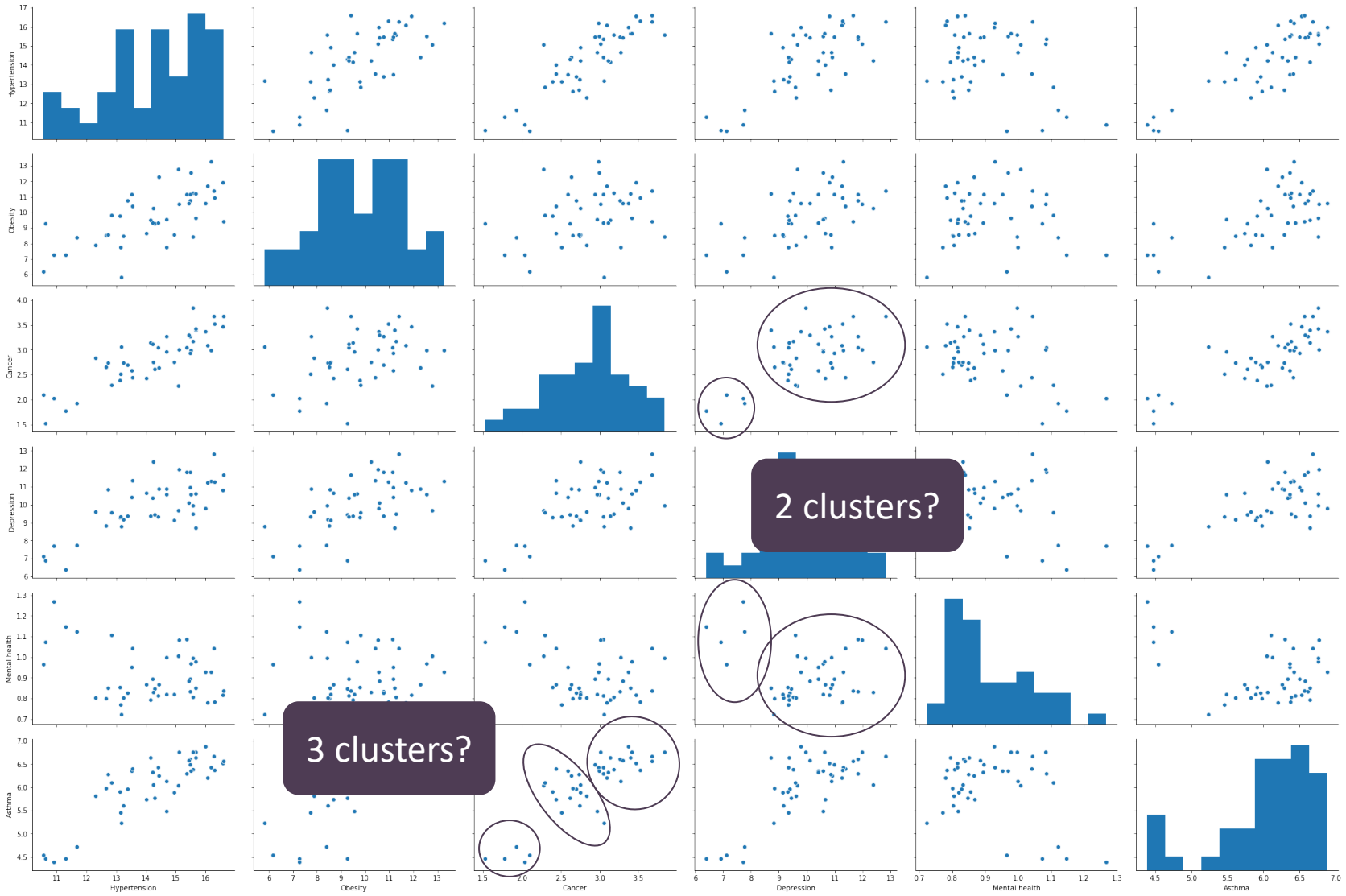
3

Clustering

- K-means
- Hierarchical

2 K-means clustering

- Deciding the number of clusters
- We have an idea from our visual checks (2 or 3)



3 K-means code

- Cycle through a range of k values to identify highest silhouette score and pick this cluster.
- Automate allocation of cluster labels to our data

```
for k in k_range:

    # Perform k means as before:

    km_output = sklc.KMeans(n_clusters = k, n_init = 20).fit(df_k_means_analyse)

    # Add the results to our lists:

    cluster_ids_series.append(km_output.labels_)
    cluster_sse_series.append(km_output.inertia_)
    cluster_cns_series.append(km_output.cluster_centers_)

    # Only add silhouettes if there is no error, since silhouettes cannot be calcula
    # ... there is only one cluster or where some points are in their own clusters:

    try:
        cluster_shs_series.append(sklm.silhouette_score(df_k_means_analyse,km_output
    except:
        cluster_shs_series.append(0)

# Build a dictionary of the quantities we need to report on and convert to a datafra
# Also drop rows with missing data:
report_dict = {'SSE':cluster_sse_series,'Silhouette Score':cluster_shs_series}
report_df   = pd.DataFrame(report_dict,index=range(k_max))
report_df   = report_df.dropna(how='any')

# Find and report the optimal values of k and r:

optimal_k_by_silhouette_score = report_df['Silhouette Score'].argmax()
optimal_silhouette_score      = report_df.loc[optimal_k_by_silhouette_score,'Silhoue

print('The optimal number of clusters, as determined by silhouette analysis, is ' +
print('The silhouette score for ' + str(optimal_k_by_silhouette_score) + " clusters

#show dataframe
#report_df

# We will add the optimal cluster ids to the main dataframe:

optimal_cluster_ids = cluster_ids_series[optimal_k_by_silhouette_score]
df_k_means_master['optimal_kmeans_cluster_ids'] = optimal_cluster_ids
```

```
The optimal number of clusters, as determined by silhouette analysis, is 2.
The silhouette score for 2 clusters is 0.5399797033855835.
```

```
/Users/antonios/opt/anaconda3/lib/python3.7/site-packages/ipykernel_launcher.py:51: Fu
tureWarning:
The current behaviour of 'Series.argmax' is deprecated, use 'idxmax'
instead.
The behavior of 'argmax' will be corrected to return the positional
maximum in the future. For now, use 'series.values.argmax' or
'np.argmax(np.array(values))' to get the position of the maximum
row.
```

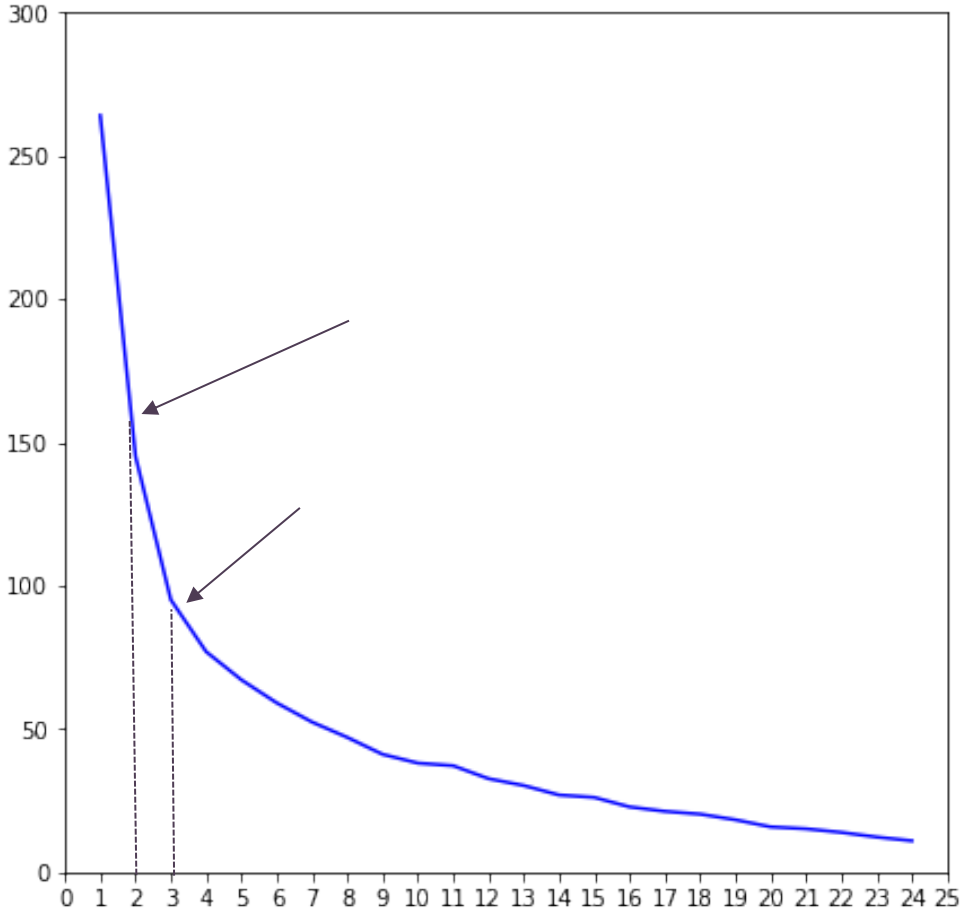
In [10]:

```
# Again, for the sake of interest, let's create an elbow plot and a silhouette plot

fignum = 20
plt.figure(fignum,figsize = (7,7))
plt.plot(report_df.index,report_df['SSE'],'b-')

#plt.gca().set_aspect('equal')
plt.gca().set_xlim([0,k_max])
plt.gca().set_xticks(range(k_max+1))
```

Code & data on GitHub [here](https://github.com/antoniosfiala/UCL_CASA_QM_NHS).
https://github.com/antoniosfiala/UCL_CASA_QM_NHS

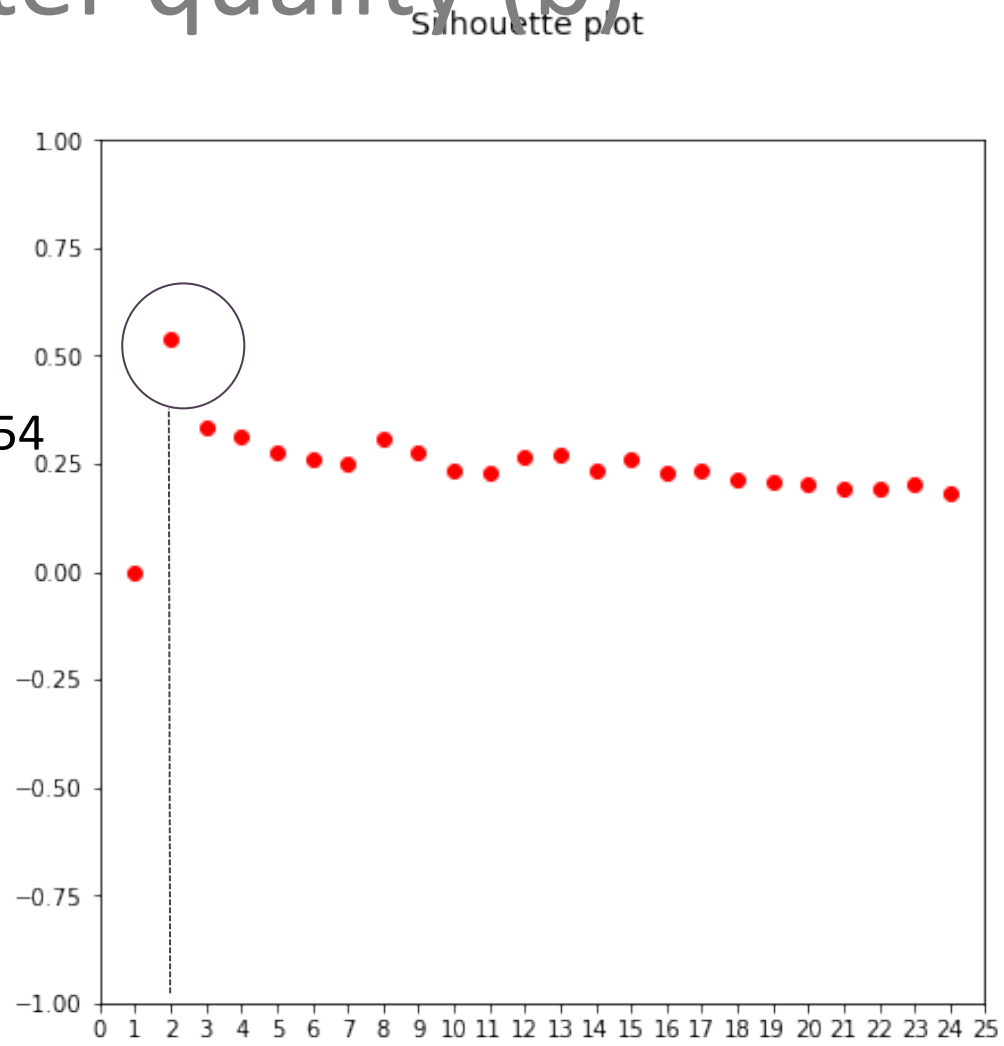
3 K-mean cluster quality (a)

- Elbow plot (SSE)
- For x value: 2 is associated with significantly higher drop in SSE than 3
- Conclusion: 2 clusters result



3 K-mean cluster quality (b)

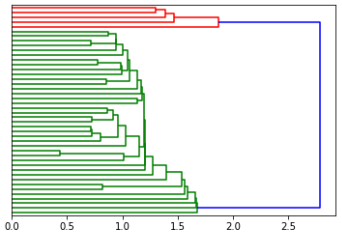
- Silhouette plot
- For x = 2 having the higher value of ~0.54 compared to x = 3 with ~0.33
- Conclusion: having 2 clusters is verified



3 Hierarchical code

In [11]:
```
#DENDROGRAM

# The following lines create and display a dendrogram, saving it as a png:

Z = spch.linkage(df_analyse, method=u_method, metric=u_metric)
spch.dendrogram(Z,orientation='right')
plt.suptitle("Dendrogram", fontsize = 14, va = "center")
plt.yticks([])
plt.figure(figsize = (7,7))
plt.show()
plt.savefig('my_dendrogram.png')

# Dendrogram (ugly axes - we'll talk about how to deal with this next week)
print()

#Z = spch.linkage(df_analyse, method=u_method, metric=u_metric)
#plt.figure(30,figsize = (10,20))
#spch.dendrogram(Z,orientation='right',color_threshold=4)

#plt.suptitle("Dendrogram", fontsize = 14, va = "center")
#plt.show()
#plt.savefig('q8_dendrogram.png')
```

Dendrogram




```
<Figure size 504x504 with 0 Axes>
```

```
report_df_h   = pd.DataFrame(report_dict_h)
report_df_h   = report_df_h.dropna(how='any')

optimal_row_by_silhouette_score_h = report_df_h['Silhouette Score'].argmax()
optimal_r_by_silhouette_score_h   = report_df_h.loc[optimal_row_by_silhouette_score_
optimal_k_by_silhouette_score_h   = report_df_h.loc[optimal_row_by_silhouette_score_
optimal_silhouette_score_h        = report_df_h.loc[optimal_row_by_silhouette_score_

print('The optimal radius, as determined by silhouette analysis, is ' + str(optimal_
print('The silhouette score for r = ' + str(optimal_r_by_silhouette_score_h) + " is
print('This corresponds to ' + str(optimal_k_by_silhouette_score_h) + ' clusters.')
```

```
The optimal radius, as determined by silhouette analysis, is 2.1052631578947367.
The silhouette score for r = 2.1052631578947367 is 0.5399797033855835.
This corresponds to 2 clusters.
```

```
/Users/antonios/opt/anaconda3/lib/python3.7/site-packages/ipykernel_launcher.py:35: Fu
tureWarning:
The current behaviour of 'Series.argmax' is deprecated, use 'idxmax'
instead.
The behavior of 'argmax' will be corrected to return the positional
maximum in the future. For now, use 'series.values.argmax' or
'np.argmax(np.array(values))' to get the position of the maximum
row.
```

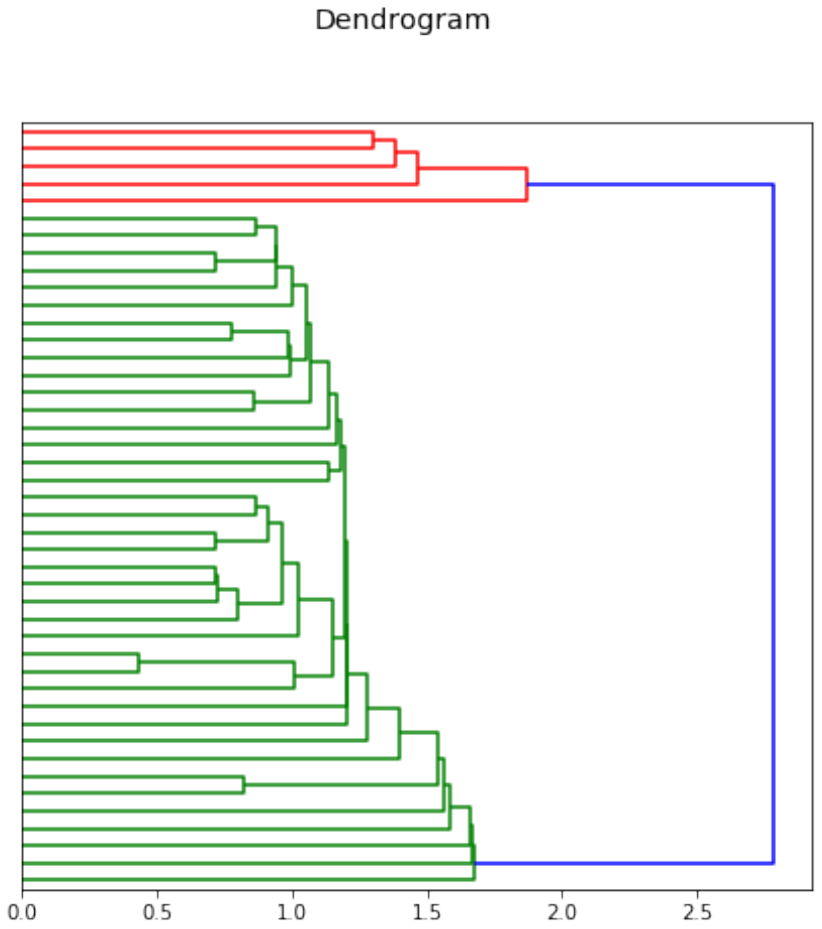
In [11]:
```
fignum = 21
plt.figure(fignum,figsize = (7,7))
plt.plot(report_df_h.index,report_df_h['Silhouette Score'],'ro')

plt.gca().set_xlim([0,int(distance_thresholds.max())])
plt.gca().set_xticks(range(int(distance_thresholds.max())+1))
plt.gca().set_ylim([-1,1])

plt.suptitle('Silhouette plot', fontsize=14, va = "center")

plt.savefig('silhouette_plot' + str(fignum) +'.png')
plt.show()
```

Silhouette plot

Code & data on GitHub here.
https://github.com/antoniosfiala/UCL_CASA_QM_NHS

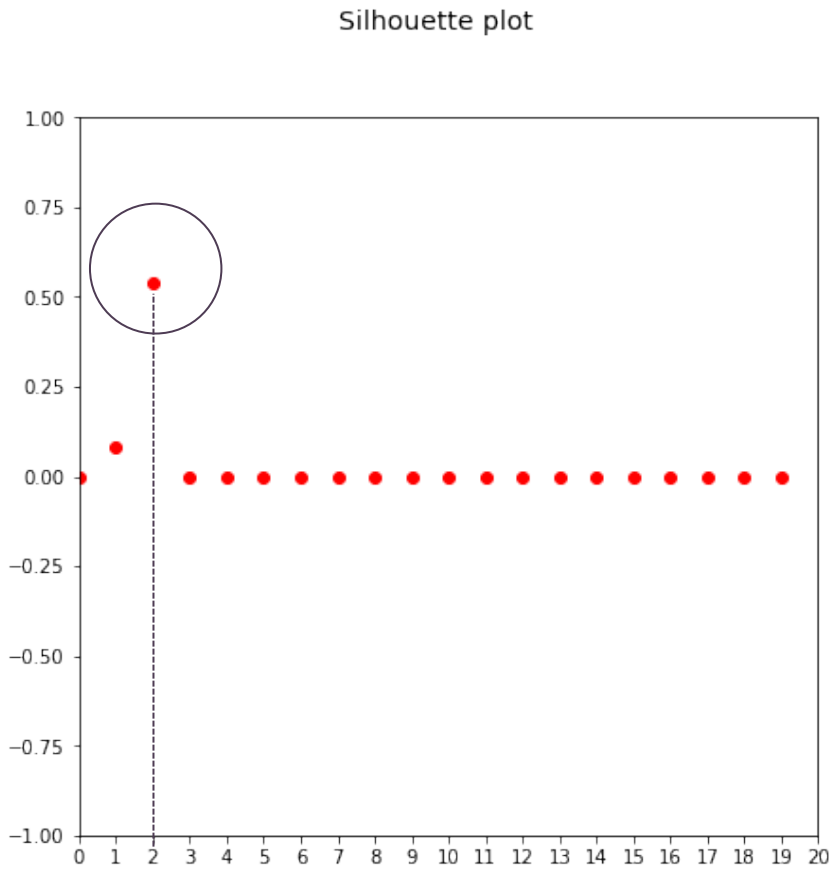
3 Hierarchical clustering quality (a)

- Dendrogram
- Radius of approximately 2 giving
  2 clusters our red and green in this case



3 Hierarchical clustering quality (b)

- Silhouette plot
- 2 clusters have a significantly higher score
- Radius of: ~2.11
- Silhouette score: ~0.54



4

Discussion & Conclusion

- Summary
- Next steps

4 Comparing method results

K-Means

Cluster 1

	Hypertension	Obesity	Cancer	Depression	Mental health	Asthma
count	39.000000	39.000000	39.000000	39.000000	39.000000	39.000000
mean	14.659963	10.067193	2.977819	10.370856	0.886981	6.235037
std	1.227229	1.592909	0.399839	1.066218	0.099210	0.401956
min	12.306925	5.820715	2.273070	8.725139	0.722217	5.237761
25%	13.532951	8.970802	2.676534	9.407543	0.815397	5.969850
50%	14.682241	10.246479	2.996502	10.406747	0.852009	6.329892
75%	15.579766	11.166555	3.268748	11.109240	0.960651	6.547431
max	16.600187	13.272109	3.840752	12.820154	1.106285	6.878169

Cluster 2

	Hypertension	Obesity	Cancer	Depression	Mental health	Asthma
count	5.000000	5.000000	5.000000	5.000000	5.000000	5.000000
mean	11.008604	7.667700	1.871514	7.176908	1.115640	4.519537
std	0.471690	1.189672	0.229079	0.575434	0.109855	0.124560
min	10.560932	6.165639	1.525505	6.384569	0.966410	4.389151
25%	10.616230	7.256099	1.774493	6.912117	1.073748	4.472412
50%	10.891937	7.256698	1.925375	7.118969	1.122379	4.474128
75%	11.307605	8.402993	2.033334	7.706745	1.147812	4.541747
max	11.666318	9.257070	2.098862	7.762141	1.267849	4.720247

Hierarchical

Cluster 1

	Hypertension	Obesity	Cancer	Depression	Mental health	Asthma
count	39.000000	39.000000	39.000000	39.000000	39.000000	39.000000
mean	14.659963	10.067193	2.977819	10.370856	0.886981	6.235037
std	1.227229	1.592909	0.399839	1.066218	0.099210	0.401956
min	12.306925	5.820715	2.273070	8.725139	0.722217	5.237761
25%	13.532951	8.970802	2.676534	9.407543	0.815397	5.969850
50%	14.682241	10.246479	2.996502	10.406747	0.852009	6.329892
75%	15.579766	11.166555	3.268748	11.109240	0.960651	6.547431
max	16.600187	13.272109	3.840752	12.820154	1.106285	6.878169

Cluster 2

	Hypertension	Obesity	Cancer	Depression	Mental health	Asthma
count	5.000000	5.000000	5.000000	5.000000	5.000000	5.000000
mean	11.008604	7.667700	1.871514	7.176908	1.115640	4.519537
std	0.471690	1.189672	0.229079	0.575434	0.109855	0.124560
min	10.560932	6.165639	1.525505	6.384569	0.966410	4.389151
25%	10.616230	7.256099	1.774493	6.912117	1.073748	4.472412
50%	10.891937	7.256698	1.925375	7.118969	1.122379	4.474128
75%	11.307605	8.402993	2.033334	7.706745	1.147812	4.541747
max	11.666318	9.257070	2.098862	7.762141	1.267849	4.720247

4 Cluster characteristics

Cluster 1

	Hypertension	Obesity	Cancer	Depression	Mental health	Asthma
count	5.000000	5.000000	5.000000	5.000000	5.000000	5.000000
mean	11.008604	7.667700	1.871514	7.176908	1.115640	4.519537
std	0.471690	1.189672	0.229079	0.575434	0.109855	0.124560
min	10.560932	6.165639	1.525505	6.384569	0.966410	4.389151
25%	10.616230	7.256099	1.774493	6.912117	1.073748	4.472412
50%	10.891937	7.256698	1.925375	7.118969	1.122379	4.474128
75%	11.307605	8.402993	2.033334	7.706745	1.147812	4.541747
max	11.666318	9.257070	2.098862	7.762141	1.267849	4.720247

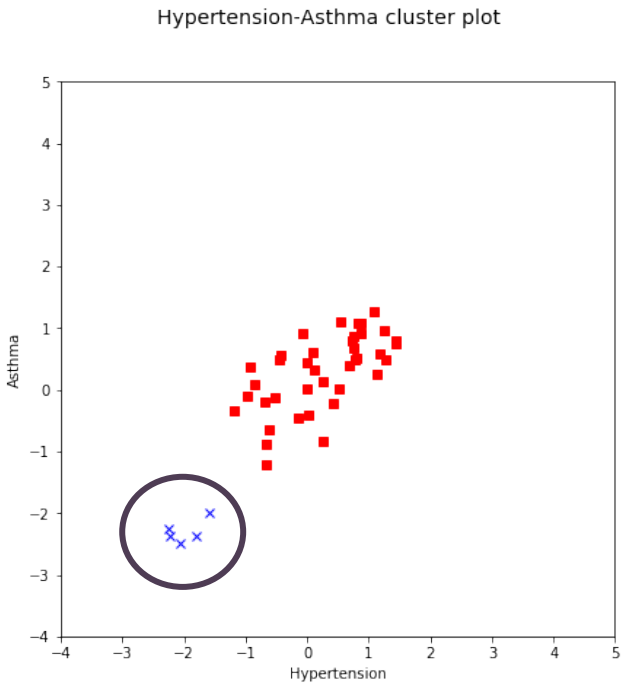
Cluster 2

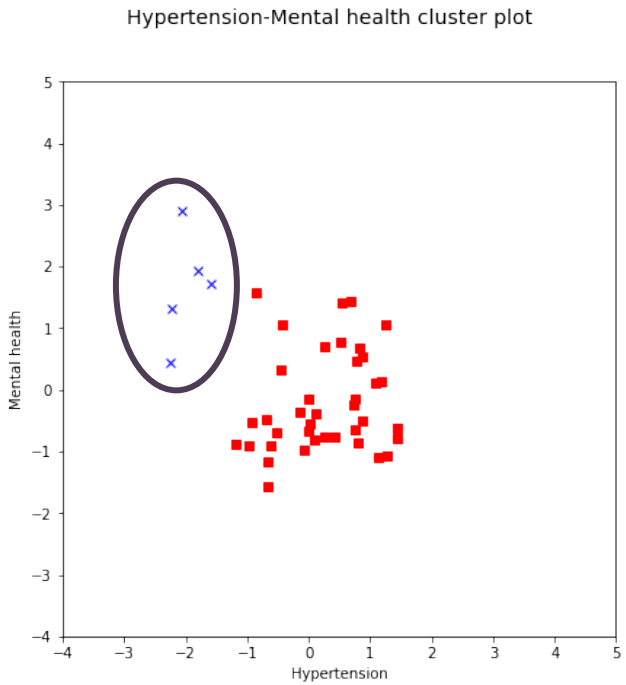
	Hypertension	Obesity	Cancer	Depression	Mental health	Asthma
count	39.000000	39.000000	39.000000	39.000000	39.000000	39.000000
mean	14.659963	10.067193	2.977819	10.370856	0.886981	6.235037
std	1.227229	1.592909	0.399839	1.066218	0.099210	0.401956
min	12.306925	5.820715	2.273070	8.725139	0.722217	5.237761
25%	13.532951	8.970802	2.676534	9.407543	0.815397	5.969850
50%	14.682241	10.246479	2.996502	10.406747	0.852009	6.329892
75%	15.579766	11.166555	3.268748	11.109240	0.960651	6.547431
max	16.600187	13.272109	3.840752	12.820154	1.106285	6.878169

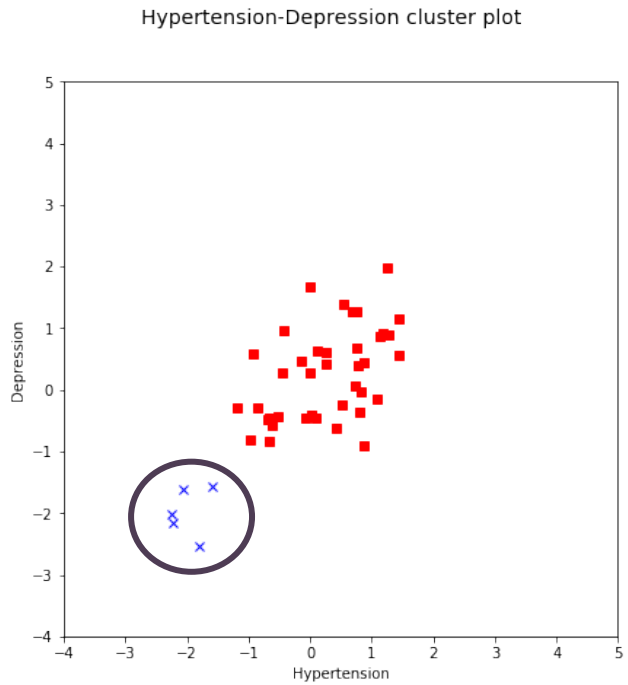
Nb of Observations: 5 & 39

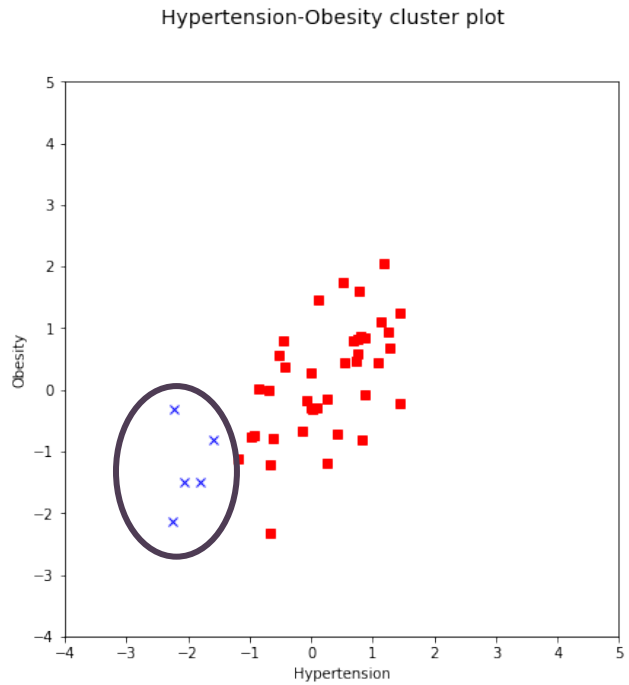
Disease level: Hypertension, Obesity, Cancer, Depression & Asthma means are higher in 2 except for mental health that is higher in 1

4 Viewing sample clusters









The STP regions are London

although most conditions have lower prevalence, mental health stands out as being higher

STPs: North West London; North Central London; North East London; South East London; South West London

4 Summary

- K means and hierarchical clustering gave us the same results
- Interpreting results:
  - Cluster 1: lower prevalence in all diseases except for mental health
  - Cluster 2: higher prevalence in all diseases except for mental health
- Cluster 1 represented 5 observations representing London
- Cluster 2 is the rest of England
- Limitations: unable to identify the specific causes of disease

4 Next steps

- Future research avenues:
  - Explore clusters separately to understand what makes London different:
    - Environment affects mental health more?
    - Reporting of mental health issues more prevalent?
    - Other reasons?
  - Add additional explanatory variables (environment, income, age...)
  - To work with a different level of data to explore hidden pattern in the clusters (e.g. CCG or practice level data)

6

Bibliography

5 Bibliography

- Tan, P.N., Steinbach, M. and Kumar, V., 2006. Cluster analysis: basic concepts and algorithms. *Introduction to data mining*, *8*, pp.487-568.
- Govender, P. and Sivakumar, V., 2019. Application of k-means and hierarchical clustering techniques for analysis of air pollution: a review (1980-2019). *Atmospheric Pollution Research*.
- Suissa S, Hemmelgarn B, Blais L, *et al*. Bronchodilators and acute cardiac death. *Am J Respir Crit Care Med*. 1996;154: 1598–602.
- TorenK,LindholmNB.Do patients with severe asthma run an increased risk from ischaemic heart disease? *Int J Epide- miol* 1996;25:617–20.
- Tattersall, M.C., Guo, M., Korcarz, C.E., Gepner, A.D., Kaufman, J.D., Liu, K.J., Barr, R.G., Donohue, K.M., McClelland, R.L., Delaney, J.A. and Stein, J.H., 2015. Asthma predicts cardiovascular disease events: the multi-ethnic study of atherosclerosis. Arteriosclerosis, thrombosis, and vascular biology, 35(6), pp.1520-1525.
- Salako, B.L. and Ajayi, S.O., 2000. Bronchial asthma: a risk factor for hypertension?. African journal of medicine and medical sciences, 29(1), pp.47-50.
- Akerman, M.J., Calacanis, C.M. and Madsen, M.K., 2004. Relationship between asthma severity and obesity. Journal of Asthma, 41(5), pp.521-526.
- NHS, 2018, Quality and Outcomes Framework, Achievement, prevalence and exceptions data - 2017-18 [PAS], electronic dataset, viewed 27 November 2019, https://digital.nhs.uk/data-and-information/publications/statistical/quality-and-outcomes-framework-achievement-prevalence-and-exceptions-data/2017-18