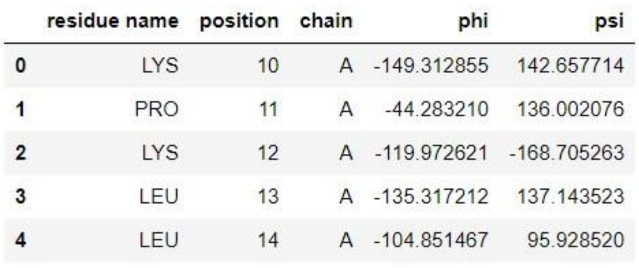
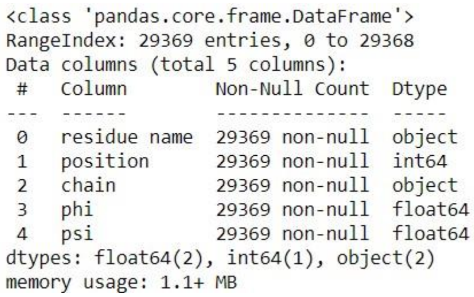
INTRODUCTION

The project is focused on the main conformation and characteristic features of the proteins and creating clusters by using the proteins’ phi and psi values via applying different clustering algorithms such as K-Means and DBSCAN. A protein chain is able to fold into its native conformation by rotation around two of the bonds in the main chain, designated f (phi) and y (psi). In order to complete the project, provided dataset named “data_all.csv” file is used and the following figure on the left demonstrates the first five rows of the dataset. The figure is on the right-hand side indicates the data types, counts and the columns names. Hence, there is no need to casting data types or dealing with the missing value problem.

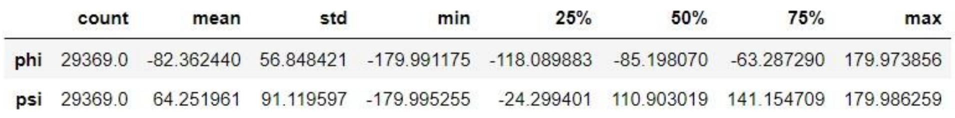


	residue name	position	chain	phi	psi
0	LYS	10	A	-149.312855	142.657714
1	PRO	11	A	-44.283210	136.002076
2	LYS	12	A	-119.972621	-168.705263
3	LEU	13	A	-135.317212	137.143523
4	LEU	14	A	-104.851467	95.928520



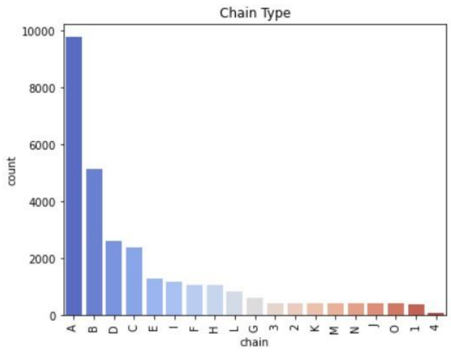
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 29369 entries, 0 to 29368
Data columns (total 5 columns):
 #   Column        Non-Null Count  Dtype
---  ------        --------------  -----
 0   residue name  29369 non-null  object
 1   position      29369 non-null  int64
 2   chain         29369 non-null  object
 3   phi           29369 non-null  float64
 4   psi           29369 non-null  float64
dtypes: float64(2), int64(1), object(2)
memory usage: 1.1+ MB
```

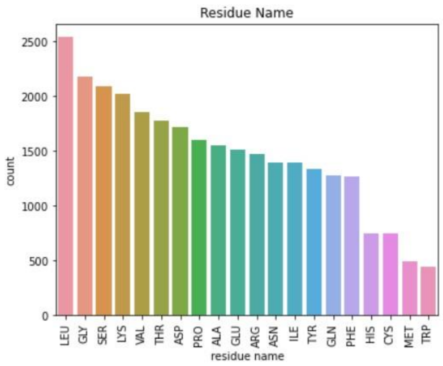
Moreover, the next figure shows the statistical description of the numeric features of the dataset. The difference between median (50%) and mean values are pretty close both for phi and psi values but in the following phase of the project and as one of the main requirement of every clustering project, different scaling approaches have applied to reduce the impacts of outliers.



	count	mean	std	min	25%	50%	75%	max
phi	29369.0	-82.362440	56.848421	-179.991175	-118.089883	-85.198070	-63.287290	179.973856
psi	29369.0	64.251961	91.119597	-179.995255	-24.299401	110.903019	141.154709	179.986259

The dataset consists of both numerical and categorical features as shown and explained above. The following figure shows the distributions of chain type and residue name features of the dataset. According to the figure, chain A and residue name LEU appeared most frequent in the dataset.

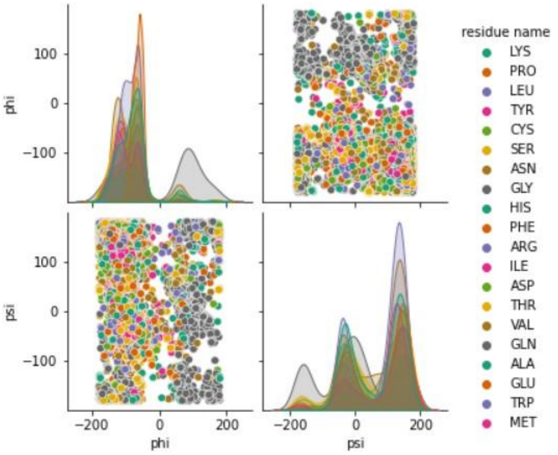


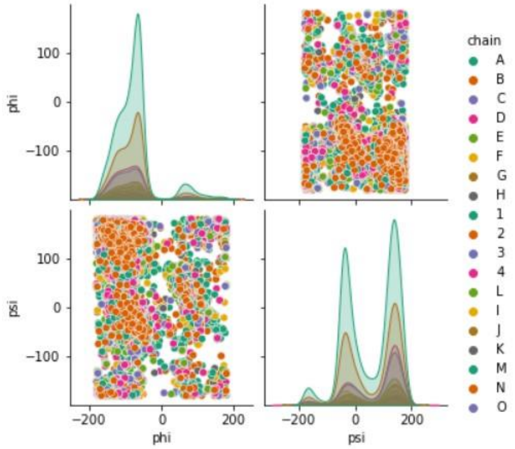


Question 1 - Show the distribution of phi and psi combinations using:

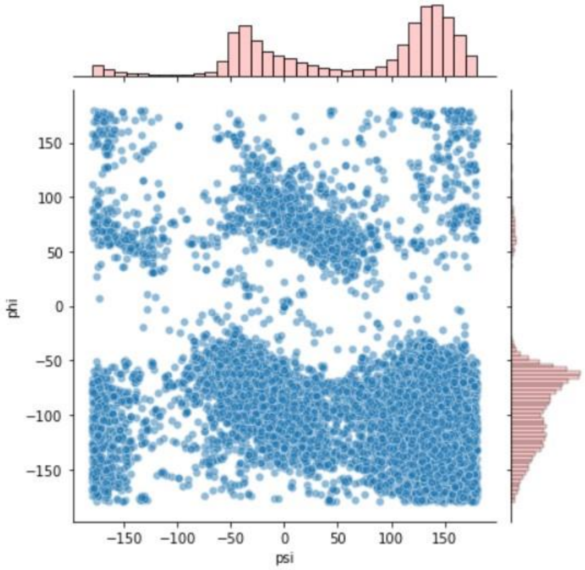
- **Scatter Plot**

Three different scatter plots have been created to reveal the distributions of phi and psi features by using other categorical features are given in the dataset by completing “hue” parameter of the plots. By looking at the next two scatter plots, there aren’t any clear clusters that can be seen before applying any clustering algorithms and distributions are not looking similar to each other in terms of residue name feature. Additionally, the figure is on the right the distribution of the psi-phi features by using chain feature looks more similar than the figure which have been created by using residue name feature.



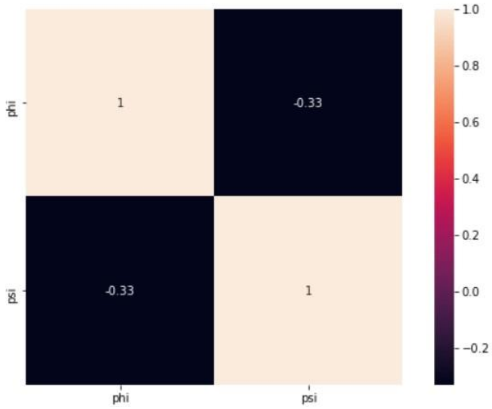


The following scatter plot has been created by only using phi and psi features without adding additional features like chain or residue name. Both distributions of psi and phi features are not normally distributed so in the light of this information, scaling is vital in the next phases of the project.



- **Heat Map**

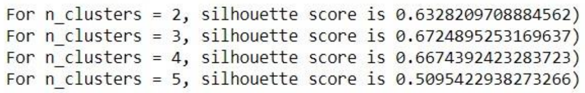
  Correlation matrices are essential and correlation heat maps are created in order to show the impacts of the numeric features on themselves. According to the figure, there is moderate/weak negative correlation which is -0.33 between the phi and psi features.



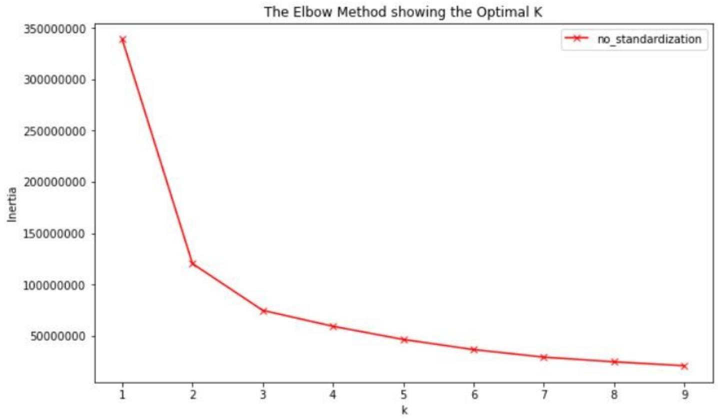
Question 2 - Use the K-means clustering method to cluster the phi and psi angle combinations in the data file.

a. Experiment with different values of K. Suggest an appropriate value of K for this task and motivate this choice.

In order to determine the number of clusters, inertia is calculated for different K values from 1 to 9 with K-Means algorithm. Besides, Silhouette Coefficient is calculated for K values from 2 to 6. Silhouette Coefficient for a set of samples can be found by taking the average of Silhouette Coefficient for each sample. According to Elbow method and Silhouette Coefficient values, the number of clusters is determined as 4 for K-Means algorithm and at the beginning, no standardization method applied to be able to see the changes of normalization techniques which will be applied in the part d of the question. The silhouette score of 1 means that the clusters are very dense and nicely separated, the score of 0 means that clusters are overlapping and the score of less than 0 means that data belonging to clusters may be wrong/incorrect.

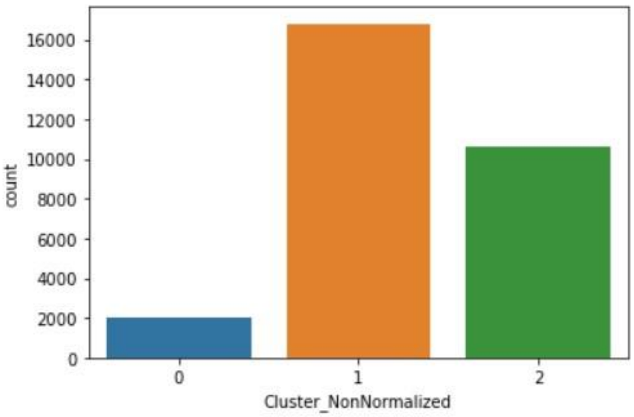


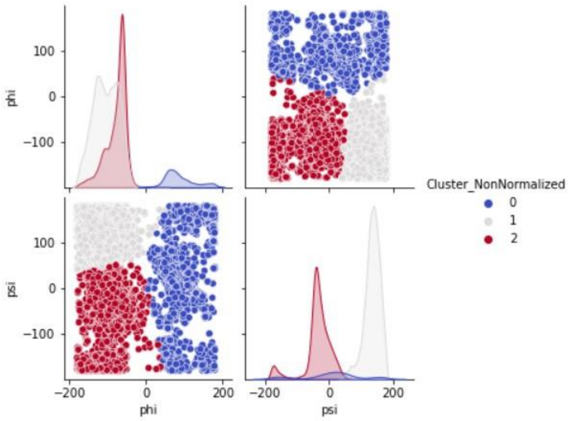
```
For n_clusters = 2, silhouette score is 0.6328209708884562)
For n_clusters = 3, silhouette score is 0.6724895253169637)
For n_clusters = 4, silhouette score is 0.6674392423283723)
For n_clusters = 5, silhouette score is 0.5095422938273266)
```



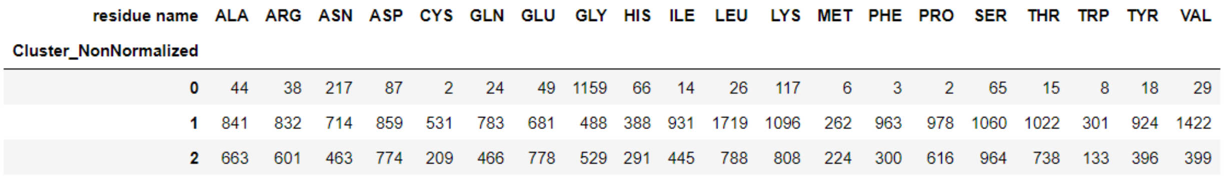
b and c. Validate the clusters that are found with the chosen value of K. Do the clusters found in part (a) seem reasonable?

Three clusters have been created and a new column called “Cluster_NonNormalized” which shows the cluster label for each sample added in the dataset. In terms of using that column, cluster distributions for Cluster 1 and Cluster 2 seem reasonable but there are fewer samples clustered in Cluster 0 as shown in the following figure on the left. Therefore, the figure is on the left demonstrates the clusters that are separated each other and its distribution.





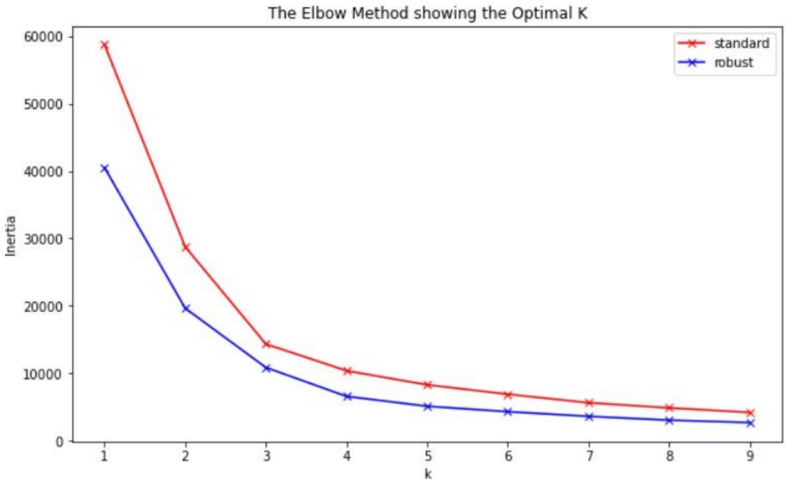
Lastly, the next table shows the cluster distributions for some of the residue name and how they are clustered:



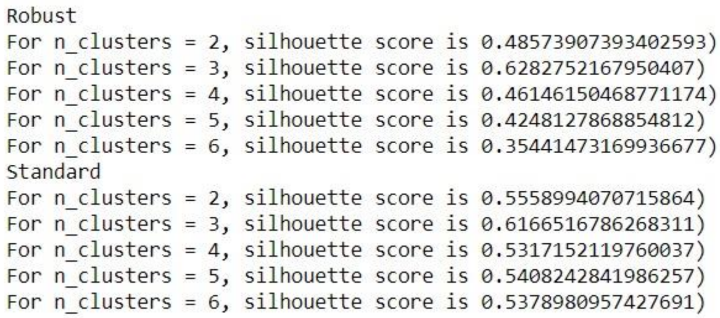
residue name	ALA	ARG	ASN	ASP	CYS	GLN	GLU	GLY	HIS	ILE	LEU	LYS	MET	PHE	PRO	SER	THR	TRP	TYR	VAL
Cluster_NonNormalized																				
0	44	38	217	87	2	24	49	1159	66	14	26	117	6	3	2	65	15	8	18	29
1	841	832	714	859	531	783	681	488	388	931	1719	1096	262	963	978	1060	1022	301	924	1422
2	663	601	463	774	209	466	778	529	291	445	788	808	224	300	616	964	738	133	396	399

d. Can you change the data to get better results (or the same results in a simpler way)?

First of all, two different normalization methods have been approaches to reduce the inertia. How spread out the samples within each cluster is can be measured by the "inertia". Intertia measures how far samples are from their centroids and desired results of the clusters that are not spread out the centers, so lower values of the inertia are better. In fact, K-Means aims to place the clusters in a way that minimizes the inertia. In a nutshell, in a good clustering project provides lower inertia and doesn't have too many clusters.

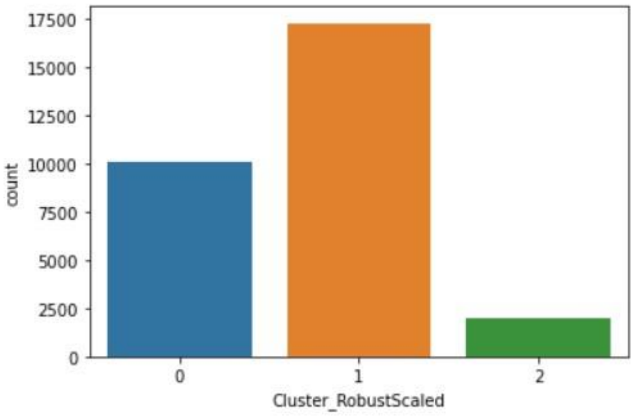


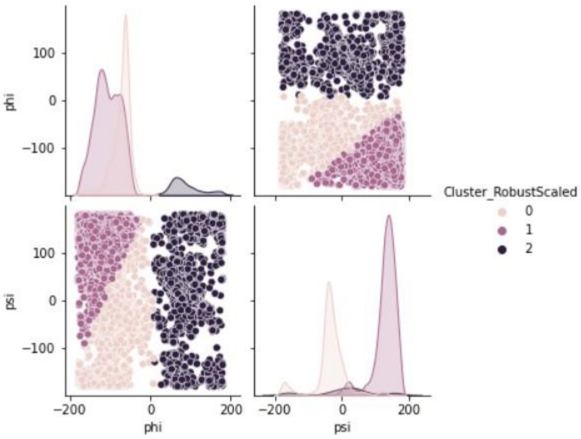
The given figure above clearly shows that robust scaling approach reached lower inertia score which is an expected and desired output from a successful clustering project. The following silhouette score calculations both for robust and standard scaler demonstrate that choosing K as 3 is the most appropriate approach for suitable number of clusters.



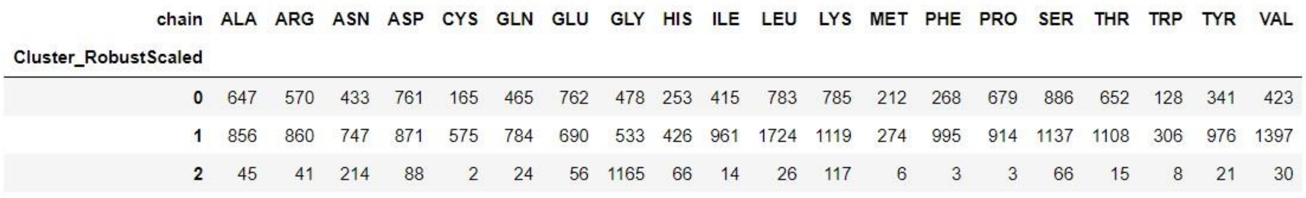
```
Robust
For n_clusters = 2, silhouette score is 0.48573907393402593)
For n_clusters = 3, silhouette score is 0.6282752167950407)
For n_clusters = 4, silhouette score is 0.46146150468771174)
For n_clusters = 5, silhouette score is 0.4248127868854812)
For n_clusters = 6, silhouette score is 0.35441473169936677)
Standard
For n_clusters = 2, silhouette score is 0.5558994070715864)
For n_clusters = 3, silhouette score is 0.6166516786268311)
For n_clusters = 4, silhouette score is 0.5317152119760037)
For n_clusters = 5, silhouette score is 0.5408242841986257)
For n_clusters = 6, silhouette score is 0.5378980957427691)
```

A similar result have been obtained in the clusters that are created by not applying normalization techniques and 3rd cluster had fewer samples than rest of the clusters as well. Besides, the figure is on the left shows the clusters that are separated each other and its distribution.



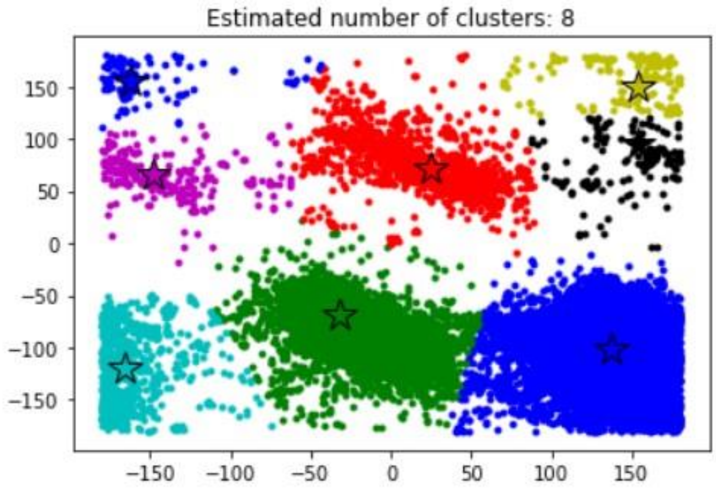


As a last part of the clustering by standardized data, the next table shows the cluster distributions for some of the residue name and how they are clustered:



chain	ALA	ARG	ASN	ASP	CYS	GLN	GLU	GLY	HIS	ILE	LEU	LYS	MET	PHE	PRO	SER	THR	TRP	TYR	VAL
Cluster_RobustScaled																				
0	647	570	433	761	165	465	762	478	253	415	783	785	212	268	679	886	652	128	341	423
1	856	860	747	871	575	784	690	533	426	961	1724	1119	274	995	914	1137	1108	306	976	1397
2	45	41	214	88	2	24	56	1165	66	14	26	117	6	3	3	66	15	8	21	30

In addition to this approach, as recommended, MeanShift clustering algorithm has used to detect the number of estimated clusters and showing the psi-phi value distributions and its clusters on a scatter plot. This approach requires less effort to cluster the samples. According to the MeanShift clustering algorithm, the number of estimated clusters is 8 and the related scatter plot can be found as below:

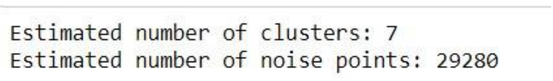


Question 3 - Use the DBSCAN method to cluster the phi and psi angle combinations in the data file.

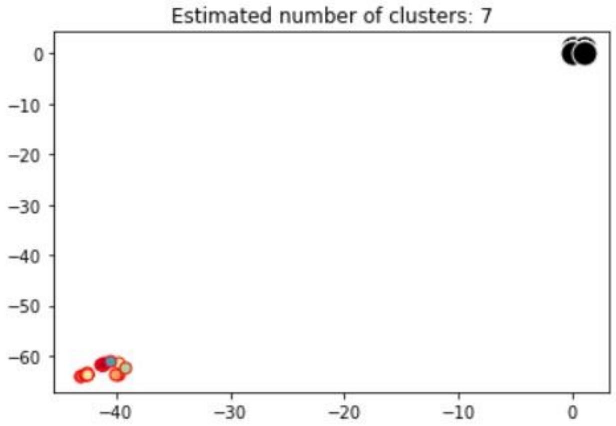
a and b. Motivate and Highlight the clusters found using DBSCAN and any outliers in a scatter plot. How many outliers are found? Plot a bar chart to show which amino acid residue types are most frequently outliers.

While creating clusters by applying DBSCAN algorithm, eps parameter which is the choice of the maximum distance between two samples belonging to the same neighborhood has been chosen as 0.3 and min_sample parameter which is the choice of the minimum number of samples in the neighborhood for a point to be considered as a core point has been chosen as 10. The parameter of DBSCAN algorithm have been changed manually to try to appropriate number of clusters like the number from 5 to 10, for instance.

According to DBSCAN clustering, estimated number of cluster has been chosen as 7 and it can be found visually in the following figure as well. Note that -1 cluster represents outliers which mean that we have 29280 data points as outliers.

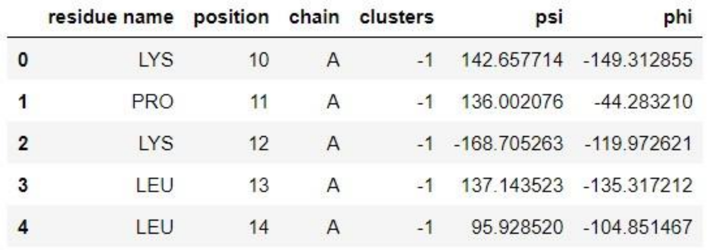


```
Estimated number of clusters: 7
Estimated number of noise points: 29280
```

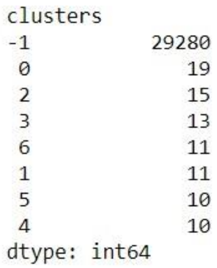


Estimated number of clusters: 7

In order to plot the bar chart to show which amino acid residue types are most frequently outliers, results are filtered only to choose the samples which is labeled as -1 which represents the outliers by DBSCAN algorithm as follows:

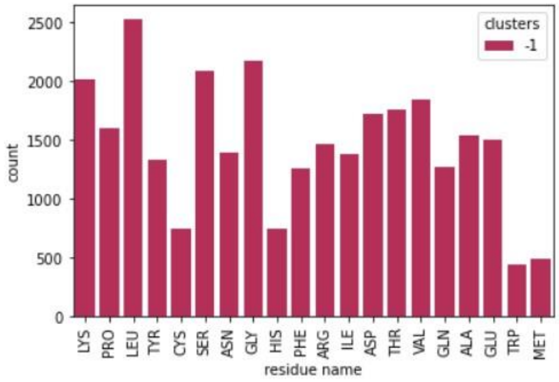


	residue name	position	chain	clusters	psi	phi
0	LYS	10	A	-1	142.657714	-149.312855
1	PRO	11	A	-1	136.002076	-44.283210
2	LYS	12	A	-1	-168.705263	-119.972621
3	LEU	13	A	-1	137.143523	-135.317212
4	LEU	14	A	-1	95.928520	-104.851467



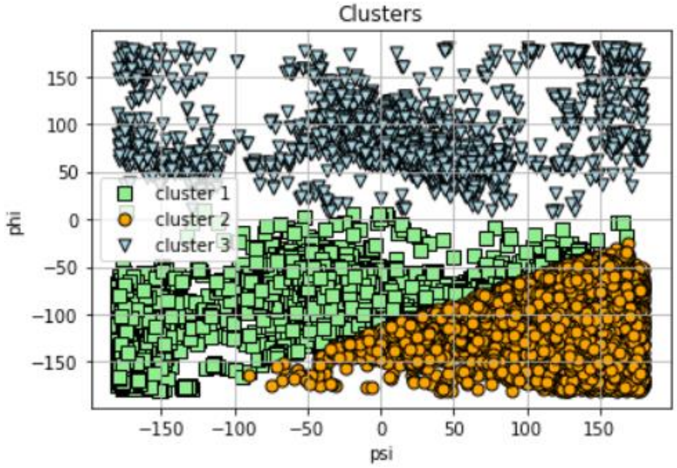
```
clusters
-1          29280
 0             19
 2             15
 3             13
 6             11
 1             11
 5             10
 4             10
dtype: int64
```

According to the results, LEU amino acid residue has the highest number of outliers as shown in the following figure:



c. Compare the clusters found by DBSCAN with those found using K-means.

The figure shows the data points and their clusters after applying K-Means algorithm as follows:



According to the results of Silhouette score and K-Means' elbow point, K has been chosen as 3 whereas DBSCAN created 7 clusters in terms of eps and min_sample parameters. So DBSCAN creates 4 more clusters than K-Means algorithm while clustering the given phi-psi features.

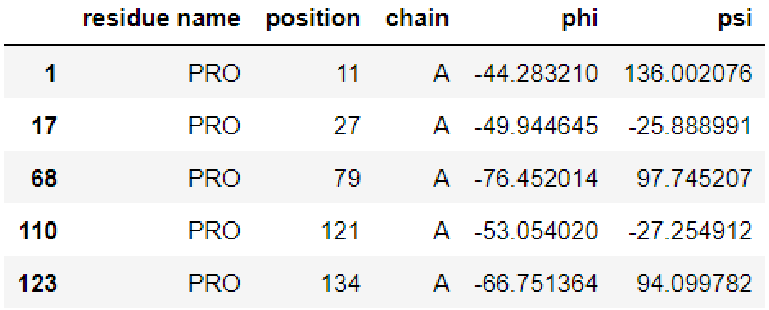
d. Discuss whether the clusters found using DBSCAN are robust to small changes in the minimum number of samples in the neighborhood for a point to be considered as a core point, and/or the choice of the maximum distance between two samples belonging to the same neighborhood ("eps" or "epsilon").

DBSCAN clustering algorithm consists of many different parameters, however, only eps and min_sample parameters have been used. While applying DBSCAN to all rows and filtering amino acids named "PRO" and "GLY", different parameters are used and that changes effected the results of the clusters. For instance, min_sample parameter is vital to decide minimum number of samples in the neighborhood really have impacts on the number of clusters and even making small changes like min_sample from 5 to 6 showed differences on the number of clusters.

Question 4 - The data file can be stratified by amino acid residue type

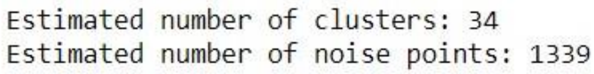
a. Use DBSCAN to cluster the data that have residue type PRO. Investigate how the clusters found for amino acid residues of type PRO differ from the general clusters

In order to provide to investigate how the clusters that have been found for amino acid's residue name is "PRO", the dataset is filtered as shown below:

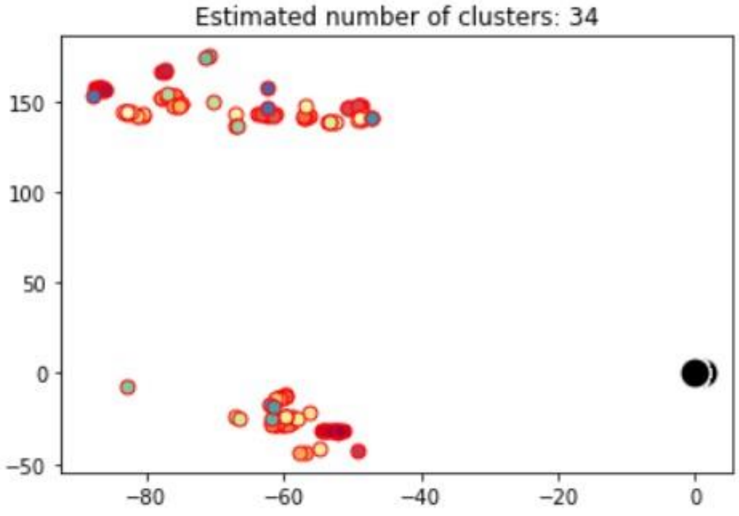


	residue name	position	chain	phi	psi
1	PRO	11	A	-44.283210	136.002076
17	PRO	27	A	-49.944645	-25.888991
68	PRO	79	A	-76.452014	97.745207
110	PRO	121	A	-53.054020	-27.254912
123	PRO	134	A	-66.751364	94.099782

After filtering the related amino acids, only phi and psi features have been selected and applied DBSCAN algorithm which is eps parameter is 1 and min_samples parameter is 5 in order to cluster the psi-phi values and the result of estimated number of clusters, noise points and their visually descriptions below. Note that black point indicates the outlier cluster.

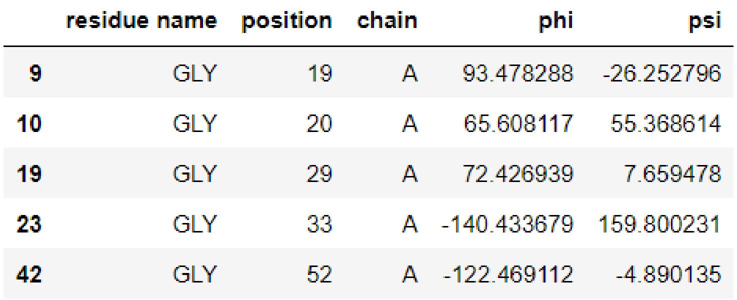


```
Estimated number of clusters: 34
Estimated number of noise points: 1339
```



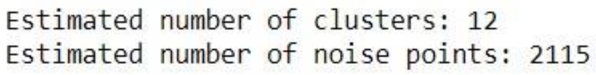
b. Now use DBSCAN to cluster the data that have residue type GLY. Investigate how the clusters found for amino acid residues of type GLY differ from the general clusters.

In order to provide to investigate how the clusters that have been found for amino acid's residue name is "GLY", the dataset is filtered as shown below:

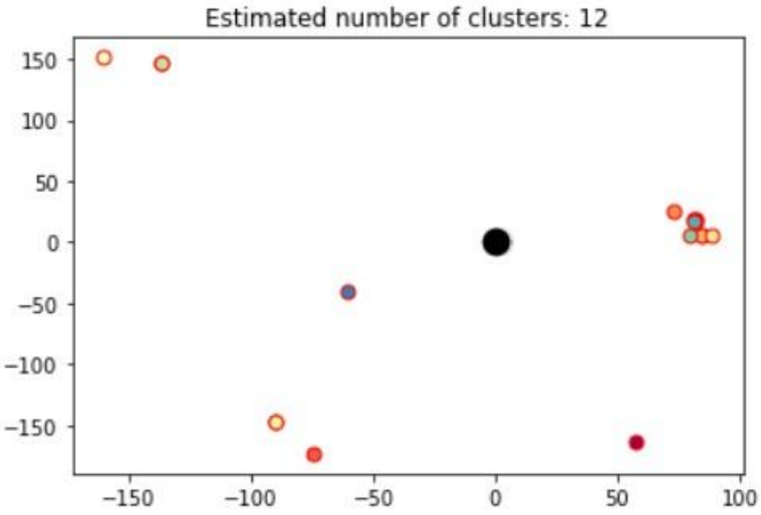


	residue name	position	chain	phi	psi
9	GLY	19	A	93.478288	-26.252796
10	GLY	20	A	65.608117	55.368614
19	GLY	29	A	72.426939	7.659478
23	GLY	33	A	-140.433679	159.800231
42	GLY	52	A	-122.469112	-4.890135

After filtering the related amino acids, only phi and psi features have been selected and applied DBSCAN algorithm which is eps parameter is 1 and min_samples parameter is 5 in order to cluster the psi-phi values and the result of estimated number of clusters, noise points and their visually descriptions below. Note that black point indicates the outlier cluster.



```
Estimated number of clusters: 12
Estimated number of noise points: 2115
```


Estimated number of clusters: 12