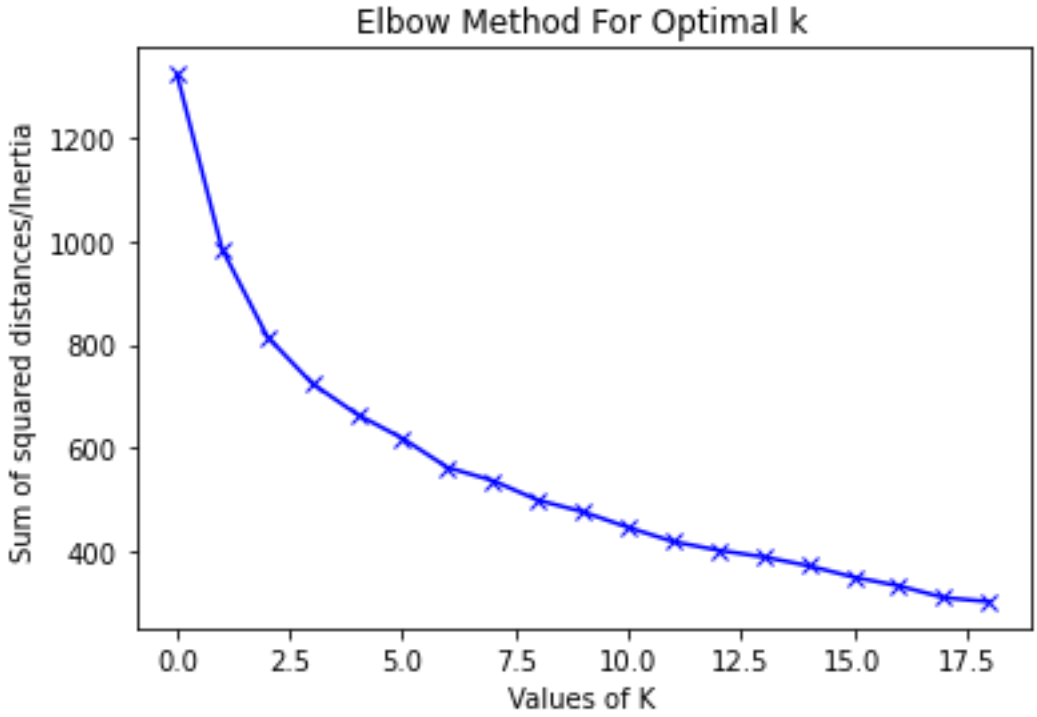
Discussions and Figures

1. Determine the number of subgroups from the dataset using attributes 3 to 205 i.e., exclude attributes 1, 2 and 206. Is this number same as number of classes presented by attribute 206? Explain and justify your findings.

A: To determine whether the number of subgroups can be achieved or not, I utilized three methods:

- Elbow method for determining the optimum number of clusters.

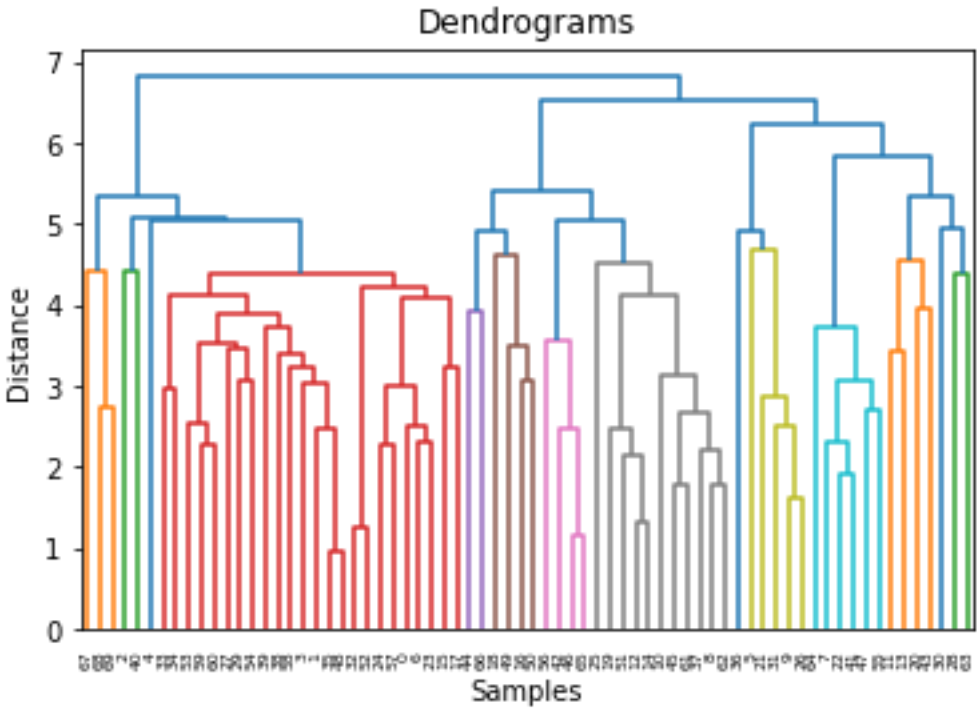
  This can not attain the original number of clusters (7 in the original dataset whereas the elbow method indicates that we should choose somewhere around 6, after which the slope becomes linear). This is encouraging result in the absence of a target variable, the KMeans method was able to look past the noise and generate meaningful insight.



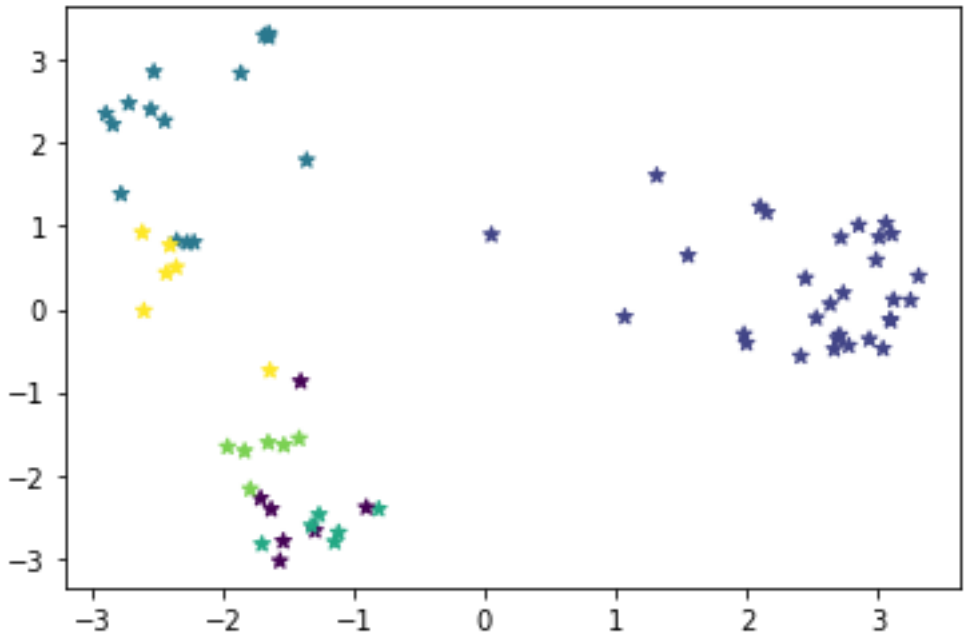
- Agglomerative Clustering to visualize cluster information.

  The agglomerative clustering was the intuitive way to go as the questions asks for subgroups, although due to imbalance in data we can not say for sure that the agglomerative clustering worked perfectly, but the performance is reasonably good.

  In the figure we can see that to obtain seven classes we can use y = 6 and the subgroups would be qual to seven which is identical to the number of classes represented by attribute 206.



- PCA scatter plot between two highest components
    - This also shows that because most of the variance was captured by the first two components, we can gain some meaningful insight in difference between clusters by just visualizing the highest two components.

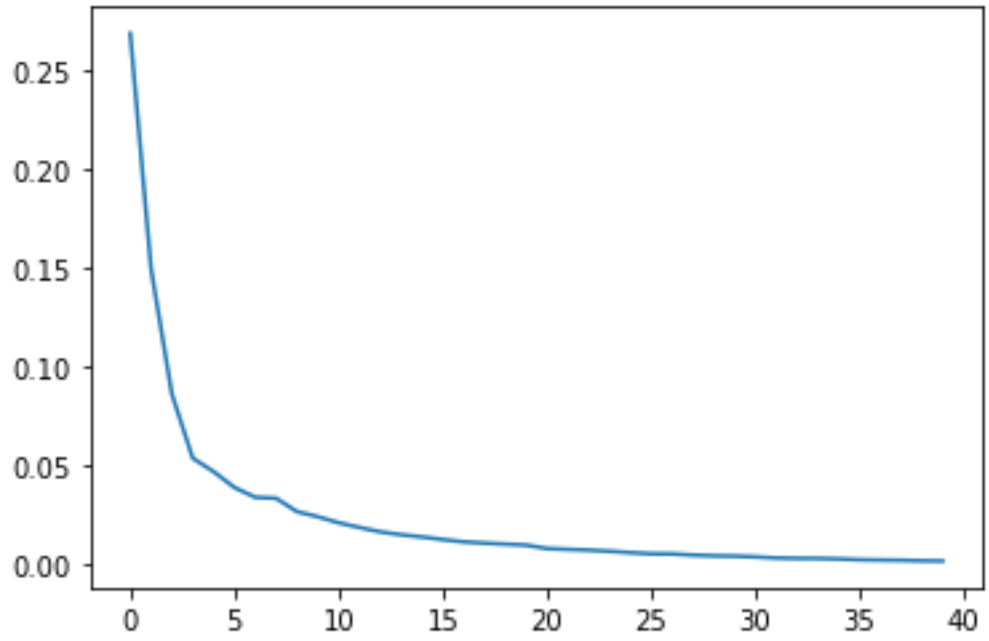


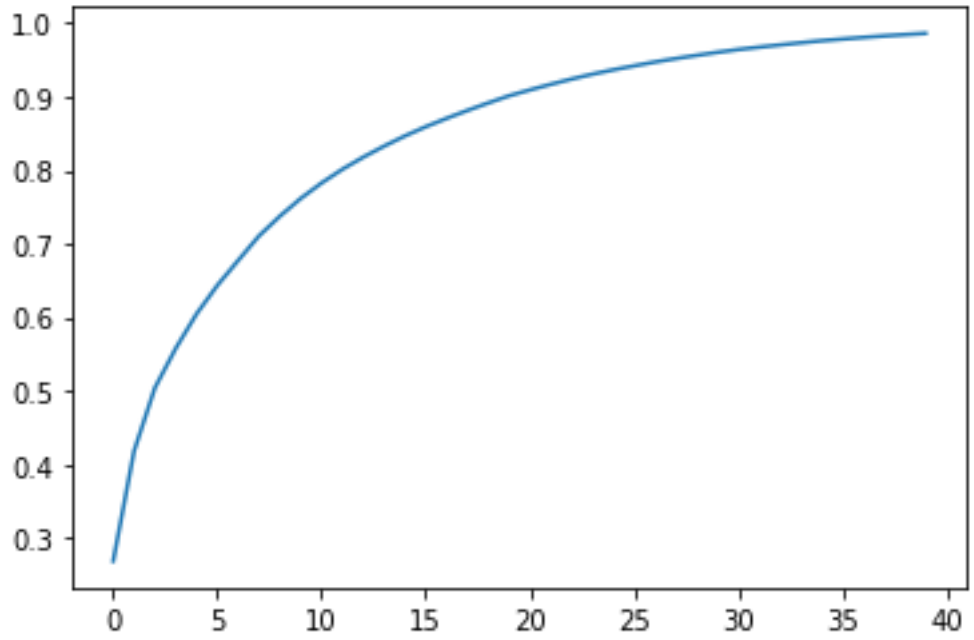
2. Is this data facing curse of dimensionality? If so, then how to solve this problem. Explain with a two-dimensional plot and report relevant loss of information.

A: Yes this data is facing the curse of dimensionality as well as class imbalance (29 samples for class6 and 1 for class3).

Classes	Age
class1	2
class2	7
class3	1
class4	12
class5	3
class6	29
class7	16

The curse of dimensionality can be validated by following two plots as we can see that after 10 components the variance explained by the data becomes linear, which denotes that PCA can be used to effectively select only those components which describe the data well (In this case 10 components constitute around 76% of the information).





3. After applying principal component analysis (PCA) on a given dataset, it was found that the percentage of variance for the first N components is X%. How is this percentage of variance computed?

The absolute percentage is computed by summing up the individual variance of all eigenvalues.
For each eigenvector, we divide the eigenvalue by the cumulative sum of eigenvalues which gives us the percentage variance explained by that eigenvector.

sum(pca.explained_variance_ratio_)

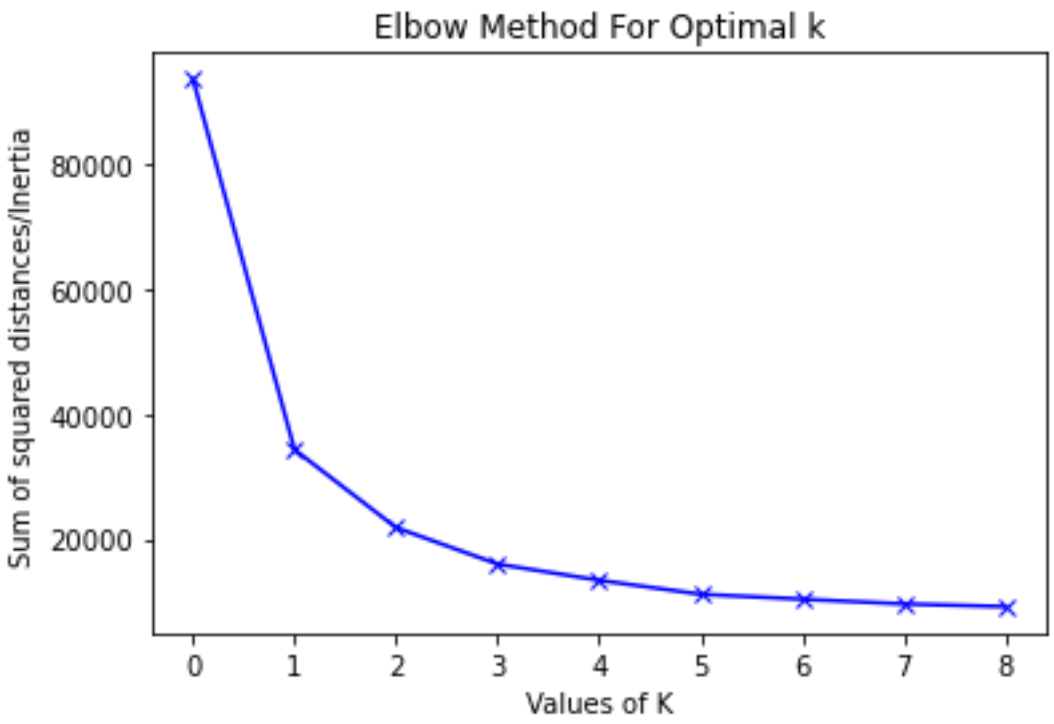
Say we get X% as the percent of variance for N components, this means that summing up N eigen values and dividing by the cumulative sum of eigenvalues gives us the percentage of variance.

4. Create a machine learning (ML) model for predicting "weight" using all features except "NObeyesdad" and report observed performance. Explain your results based on following criteria:

a. **What model have you selected for solving this problem and why?**

I have used KMeans with elbow method to solve this problem because in absence of the knowledge of target variables, elbow method gives a good idea of what should be the optimal number of clusters for clustering.

We can observe from following figure that after 2 clusters (sharp edge at 1, but 1 cluster does not give any info so going with two) the slope of the elbow curve starts flattening, this shows that the ideal number of clusters should be 2. This shows that KMeans is not able to identify the pattern in data to predict weight accurately.



b. Have you made any assumption for the target variable? If so, then why?

I am assuming that the cluster correspond to range of weights. So, if a new sample arrives it would be placed in a cluster and it's wight would be the average weight of all the samples in the cluster.
Because we do not have information about weight at the time of modelling it is impossible to assign a numerical value to the weight. This denotes the limitations of unsupervised learning in making sense of data and also highlights the strength of being able to cluster similar samples without knowing full details.

c. What have you done with text variables? Explain.

As KMeans can not handle categorical text variables, we need to encode them.

For encoding, there are two popular methods:

- One hot encoding
- LabelBinarizer

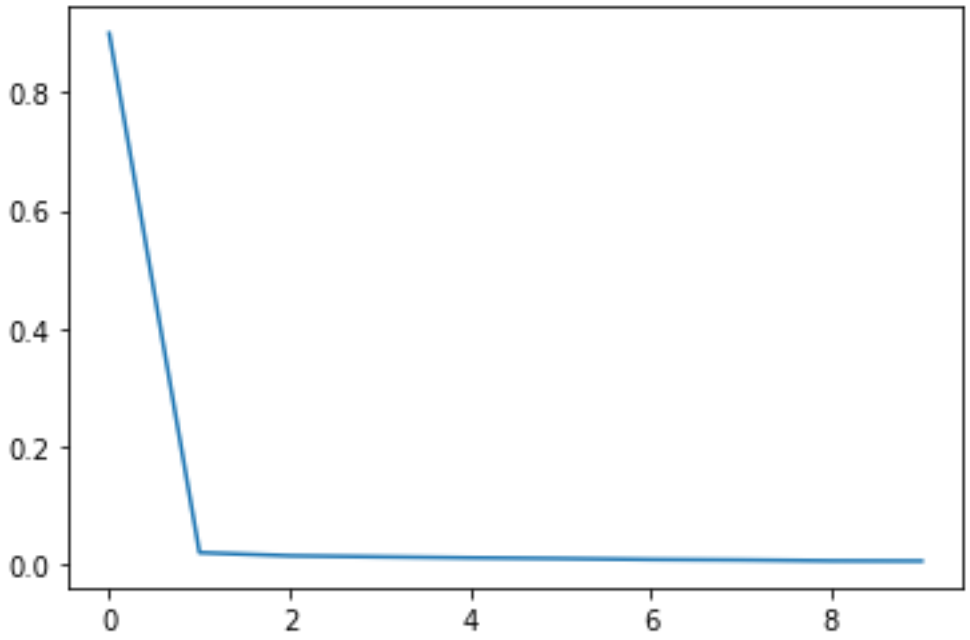
I have used one hot encoding because the properties were mutually orthogonal, and all the relevant information could be captured if we couple PCA with One hot encoder which returns a sparse matrix.

d. Have you optimised any model parameters? What is the benefit of this action?

I tried changing the number of iterations of KMeans, but the results were similar. So, no optimization was achieved.

e. **Have you applied any step for handling overfitting or underfitting issue? What is that?**

As it is an unsupervised approach the possibility of overfitting arises only when there is a dimensionality problem in data, which I validated through PCA. So after applying PCA we can assume that the model wont overfit or underfit.
Although by cross validating with ground truth data we can see that this approach fails as the **clustering criteria does not represent weight groups**.



5. Create a ML model for classifying subjects into two classes applying following constraints on above dataset.

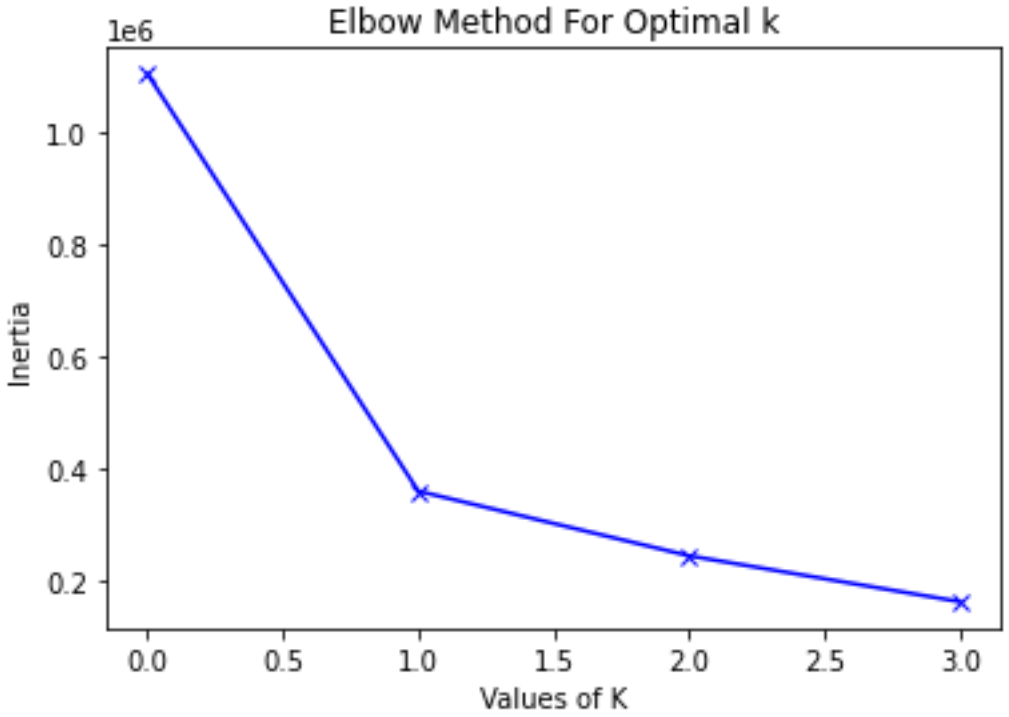
• Use "NObeyesdad" as target variable and rest of them as predictor variables.

• drop samples with value "Insufficient Weight" for "NObeyesdad"

• Group Normal Weight, Overweight Level I, and Overweight Level II into a class, and the other three labels (Obesity Type I, II, III) as the other class.

My interpretation of this problem is that we have to drop the target variable "**NObeyesdad"** and through our unsupervised approach try to recreate the target groups.

For this I have again used elbow method, in this case it works well and we can see that the suggested number of clusters is closer to the ground truth. We can say that clustering is being done to bin the weights into the respective classes.



a. Report classification performance scores. Select scores that you think best for describing the model performance with appropriate justification.

For this problem the best score would be the adjusted rand score as it takes into account ground truth as well as minimizes the chance error in rand index score (High score, false indication of good clustering performance). The adjusted rand score of 0.57 denotes that the performance is average and there is some possibility of wrong clustering of new data.

a. silhouette score for KMeans: 0.5712224623932787
b. rand score for KMeans: 0.7892477164755175
c. adjusted rand score for KMeans: 0.5784766969326518

b. Have you taken any step to check generalisability of the model? What is that and how it ensures generalisability.

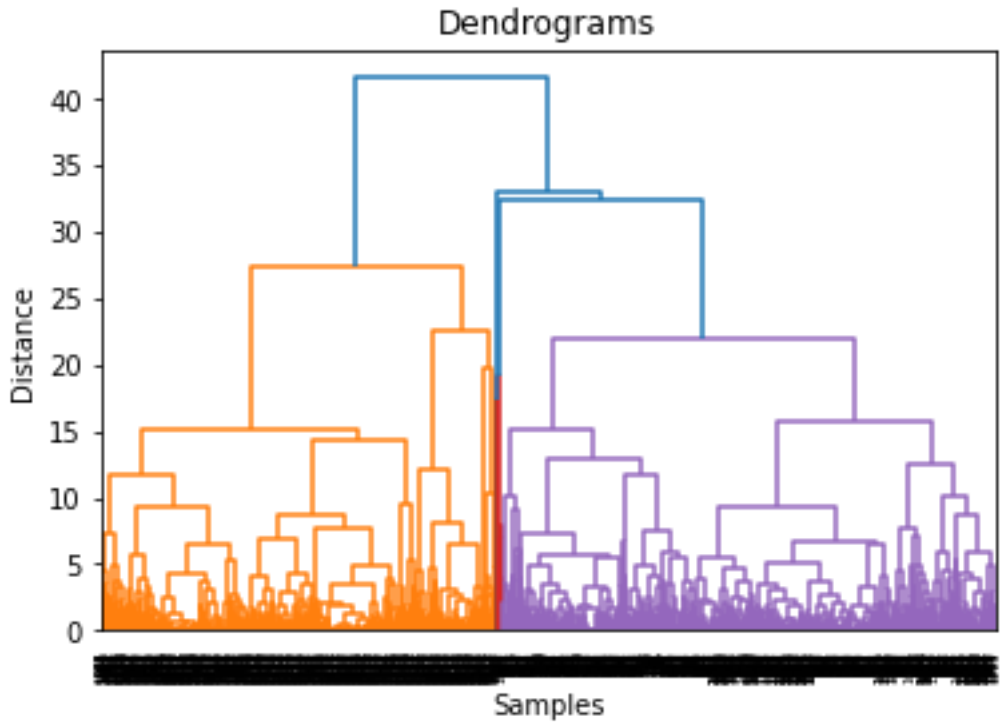
For generalisability of the model I have used unsupervised learning which is inherently generic as the model is trained without taking into the account any target variable for minimizing distance between prediction as ground truth. Also, this couple with PCA makes this model highly generalised.

c. Can you design and develop any other model for solving this problem? If so, then why have you used the reported one? Give your justification.

There are few alternate approaches that I have tried

- Kmeans with and w/o PCA [Similar result but with PCA was computationally cheaper]
- Agglomerative Clustering (Figure Below): This shows two distinct clusters.

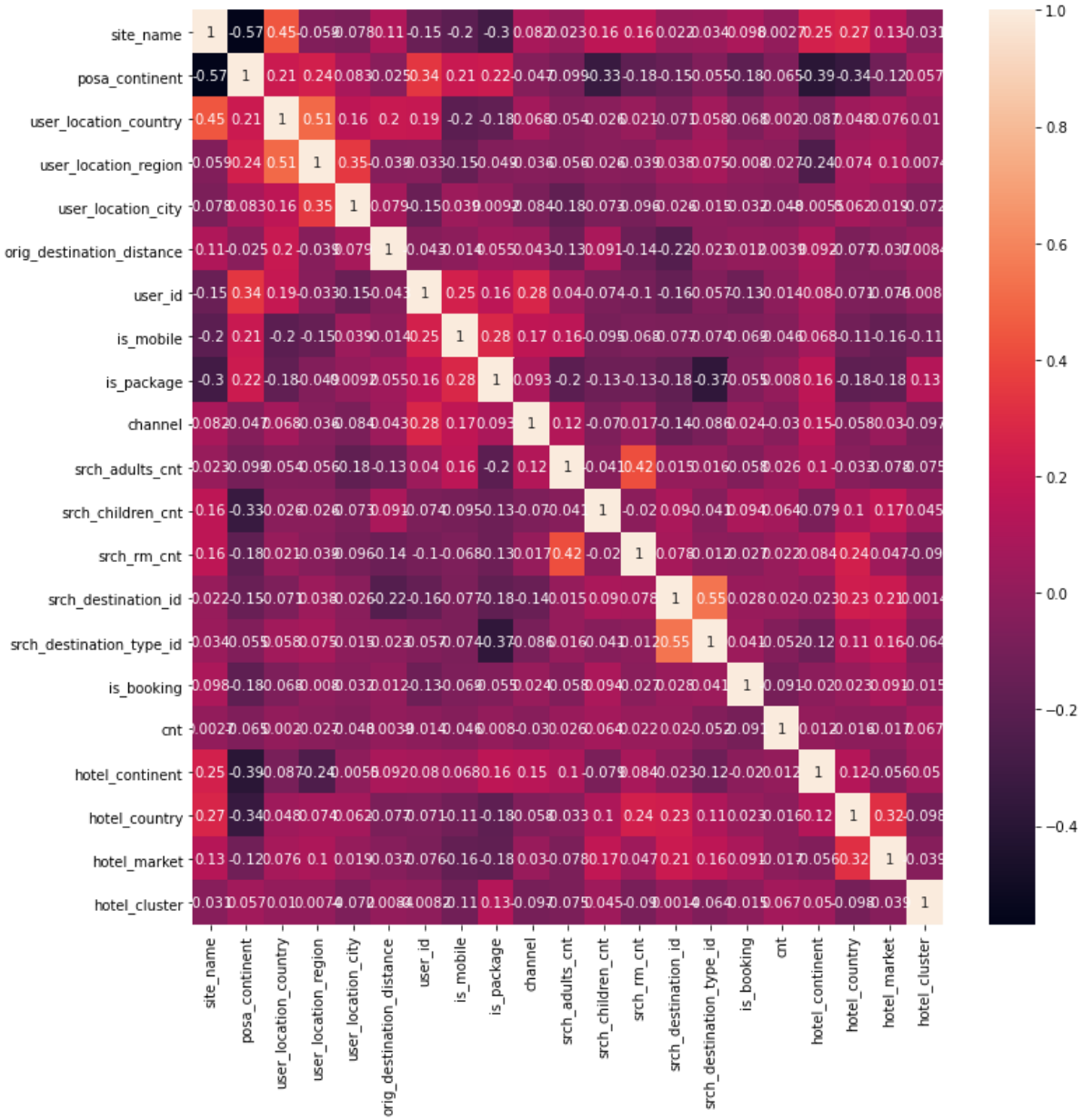
Although, my preferred approach would be to do oversampling to remove/handle the class imbalance and then go with clustering. I have not used this approach as it require generating/manipulating given dataset. Also, if target variable could be utilized then I would go with a supervised classifier (KNN, SVM etc.)



6. Suppose that a company has a number (>=500) of resorts around the globe.

- **Identify a list of features (>=5) that can be used to describe these resorts.**

To identify potential list of features to describe hotels I have calculated covariance matrix which would help In identifying the most important features that can be used to build a ML model.



- **Create a dataset (rows>=500) and explain all variables. You can generate data either synthetically or collecting from similar datasets. Submit your created dataset. In addition, please provide links in case you have collected the dataset.**

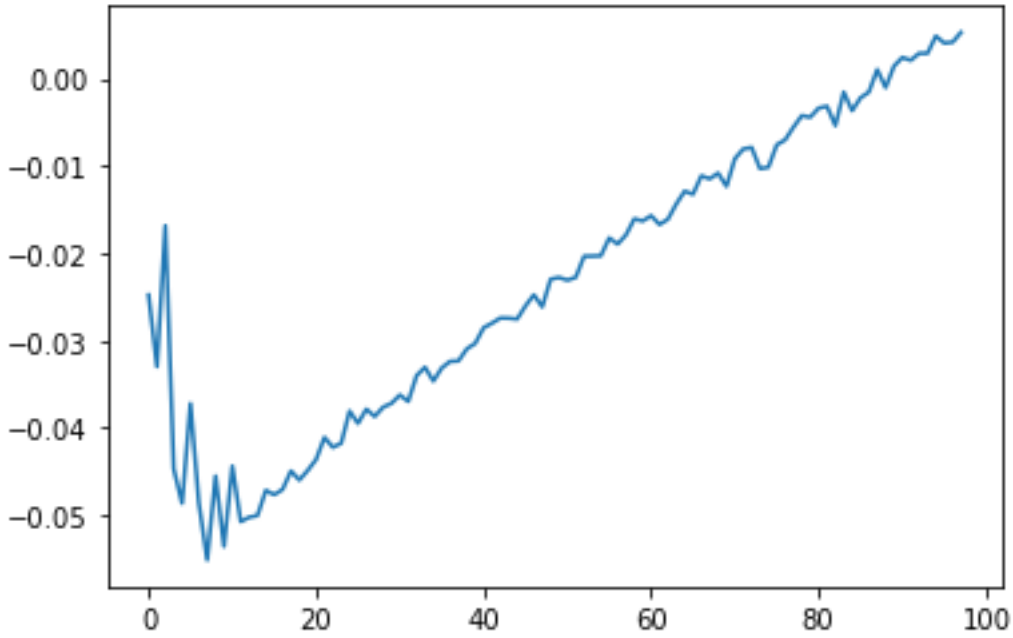
  For this problem I have taken the expedia hotels dataset as it captures real world variance and also has wide range of features. (Please ignore the destinations.csv file from the dataset as it contains extra information not relevant to this question)

Table 1: Source [1]

date_time	Timestamp
site_name	ID of the Expedia point of sale (i.e. Expedia.com, Expedia.co.uk, Expedia.co.jp, ...)
posa_continent	ID of continent associated with site_name
user_location_country	The ID of the country the customer is located
user_location_region	The ID of the region the customer is located
user_location_city	The ID of the city the customer is located
orig_destination_distance	Physical distance between a hotel and a customer at the time of search. A null means the distance could not be calculated
user_id	ID of user
is_mobile	1 when a user connected from a mobile device, 0 otherwise
is_package	1 if the click/booking was generated as a part of a package (i.e. combined with a flight), 0 otherwise
channel	ID of a marketing channel
srch_ci	Checkin date
srch_co	Checkout date
srch_adults_cnt	The number of adults specified in the hotel room
srch_children_cnt	The number of (extra occupancy) children specified in the hotel room
srch_rm_cnt	The number of hotel rooms specified in the search
srch_destination_id	ID of the destination where the hotel search was performed
srch_destination_type_id	Type of destination
hotel_continent	Hotel continent
hotel_country	Hotel country
hotel_market	Hotel market
is_booking	1 if a booking, 0 if a click
cnt	Numer of similar events in the context of the same user session
hotel_cluster	ID of a hotel cluster

- **Build a ML model that can help a customer to select appropriate set of resorts based on the season of travel. Present and describe the performance of your model.**

  For solving this problem, I have taken KMeans algorithm as we have to use unsupervised approach. To choose optimum number of clusters I came up with a method in which I have plotted Silhouette score for different number of clusters.
  We can observe from the graph that as the number of clusters increase the silhouette score also keeps increasing (ground truth 98). This shows that clustering works well, although the problem that clustering must be based on season is still unsolved as we can not infer the numerical/categorical value which a cluster denotes.



- **Why do we need a ML model for this problem?**

  We need a ML model to solve this problem because the size of data is very large which can not be handled by hand crafted rules, whereas a ML model will help in identifying hidden patterns from the dataset. In this case if season is taken as a target variable, then a supervised learning approach can yield in better results for the customer as the model would minimize the distance between ground truth “season” and predicted list of hotels based on season.

References:

[1] https://www.kaggle.com/c/expedia-hotel-recommendations/data