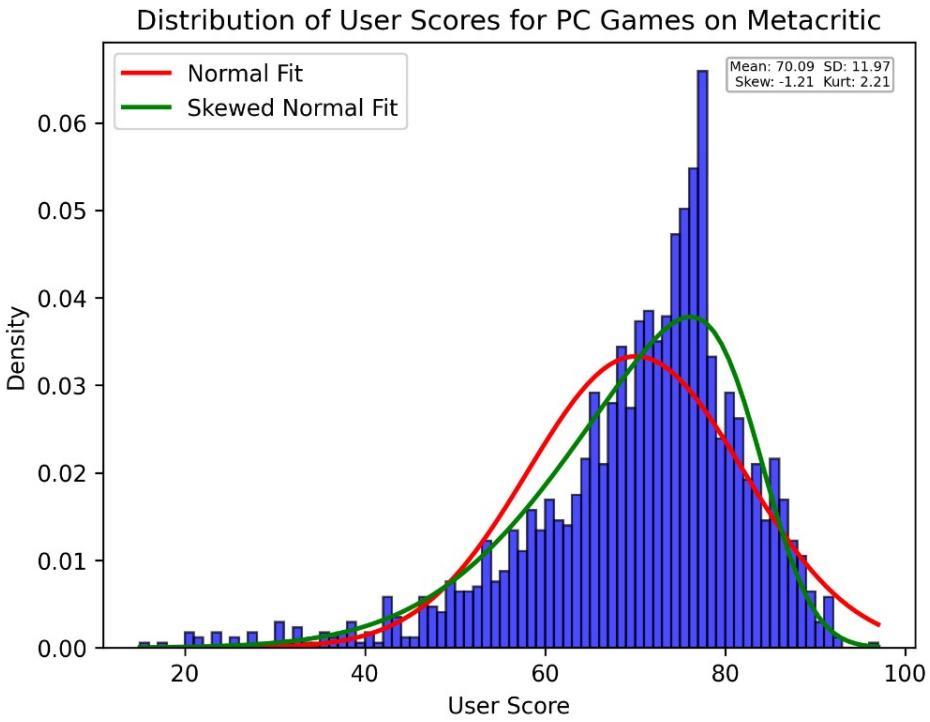
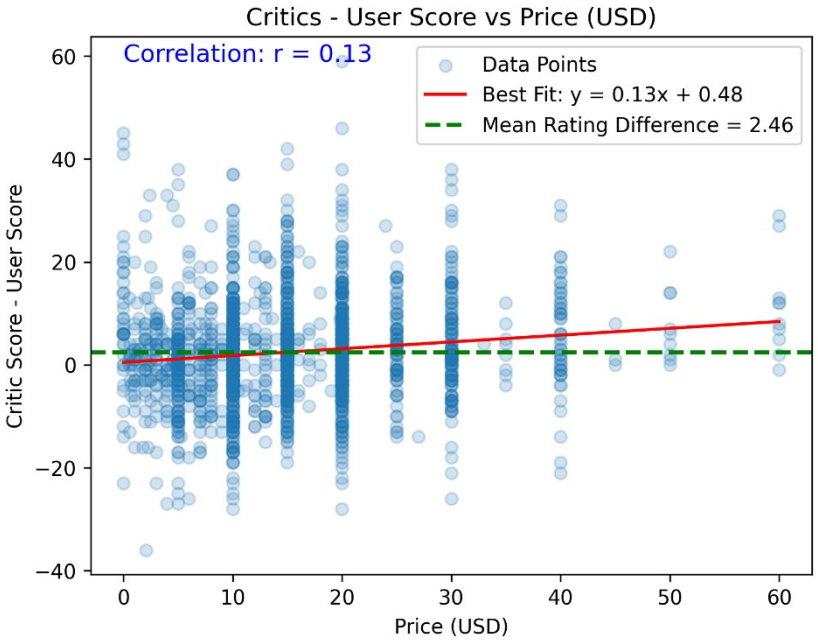
Results

Initial Data Exploration



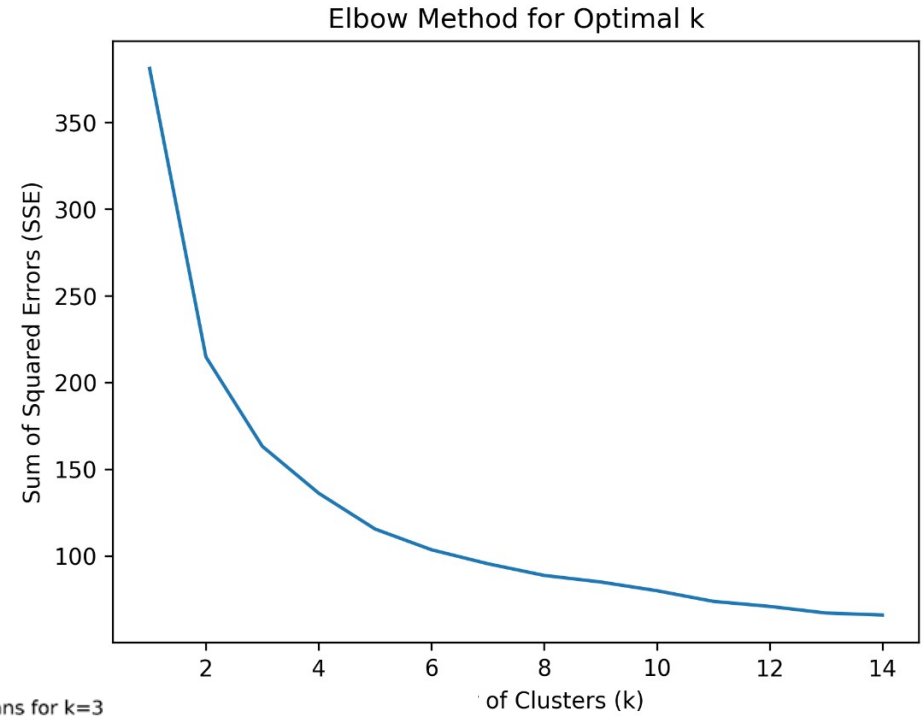
Graph showing the distribution of Metacritic User Score for PC Games, showing that the data best follows a skewed normal distribution.

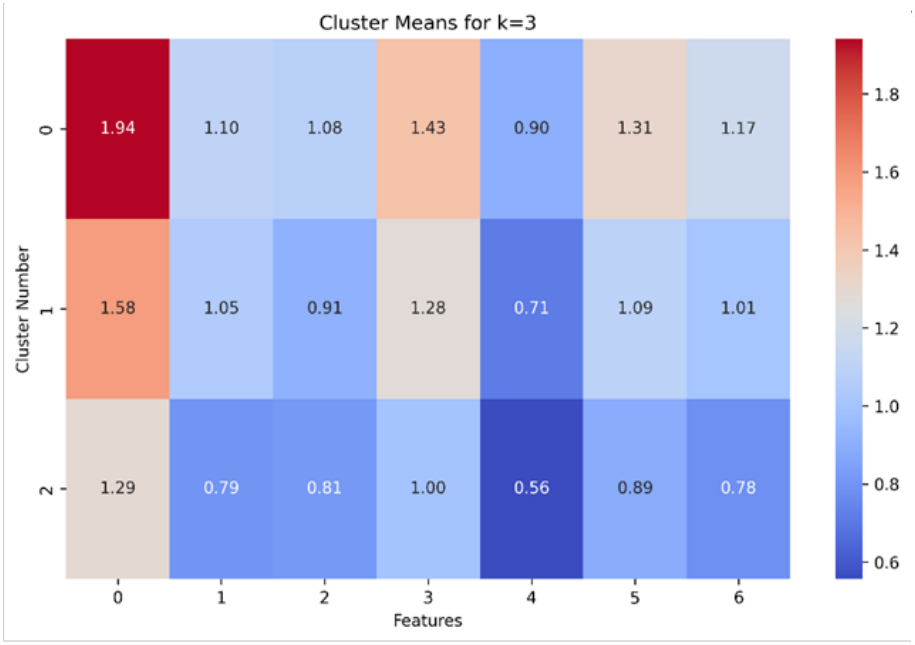


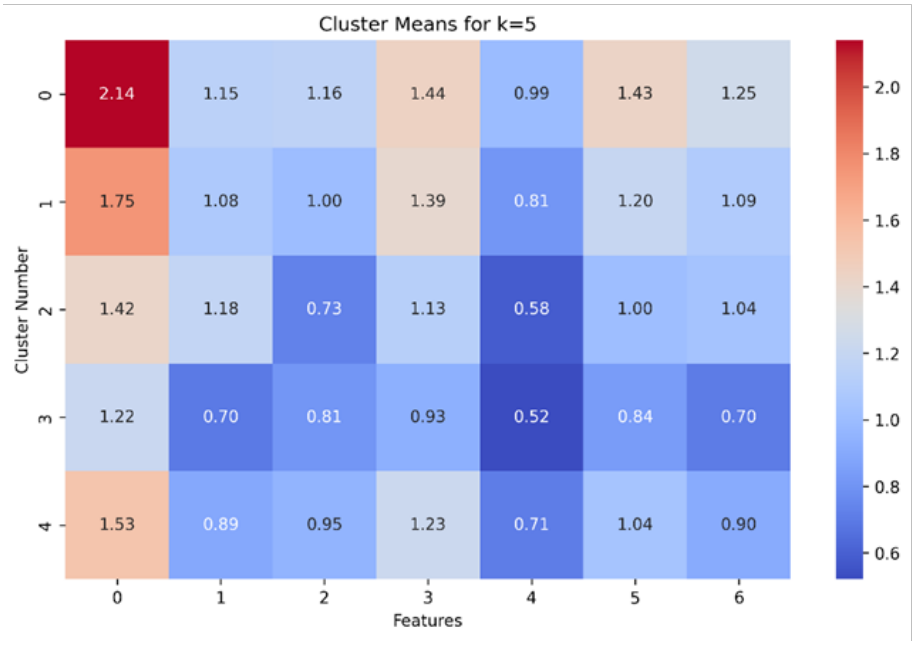
This graph shows User Score subtracted from Critic Score, plotted against price. The null hypothesis H_0 was that 'Price does not have a larger effect on user score than on critic score as price increases.' With H_1 'Price has a larger effect on user score than on critic score as price increases.' Using the Fisher Significance Testing, the probability that H_0 was true was 0.07% <1%, a level of statistical significance far smaller than my acceptance threshold so I rejected the null hypothesis and accepted H_1.

Unsupervised Learning

This graph shows how the sum of squared errors changes for varying values of k – number of clusters. From this graph I chose to look at k=3, k=5 using the elbow method to minimise in-cluster variance while preventing overfitting.



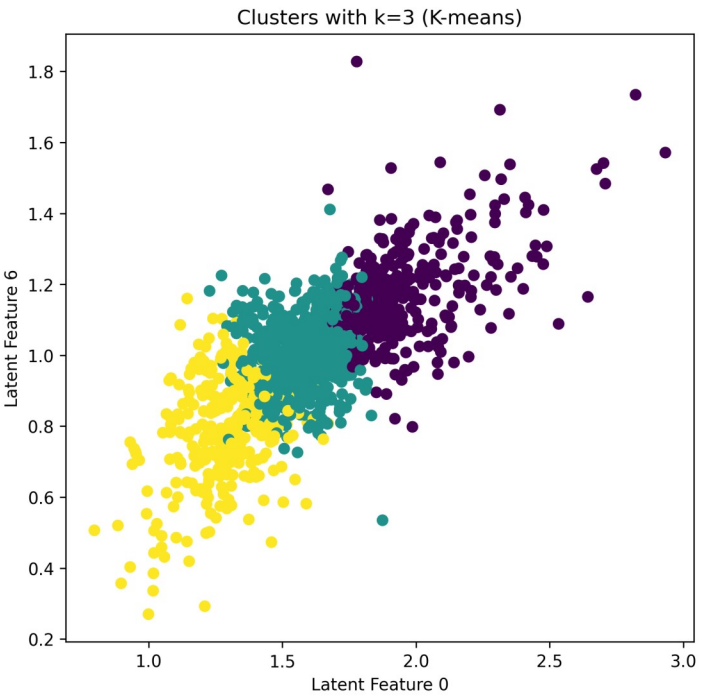


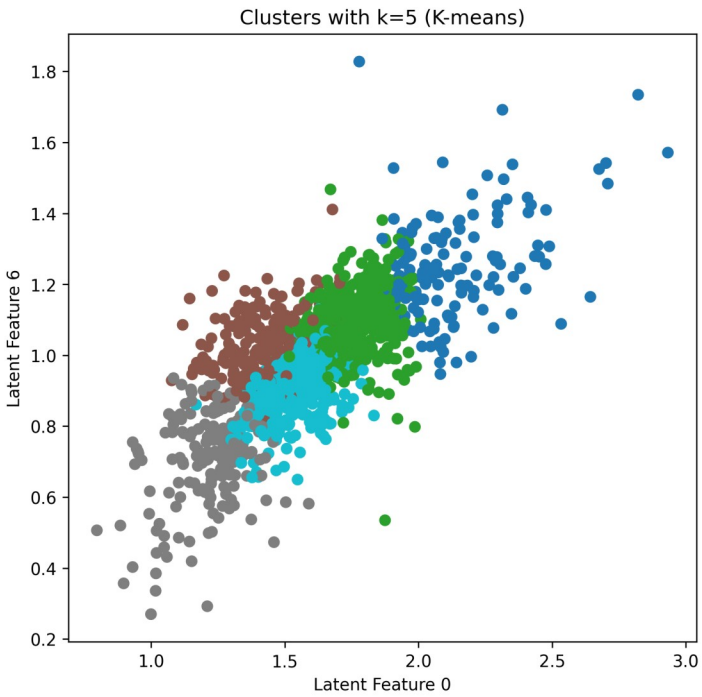


These heatmaps show the mean value of each latent feature for each cluster. From these we can see the distinguishing characteristics of each cluster in terms of the latent features.

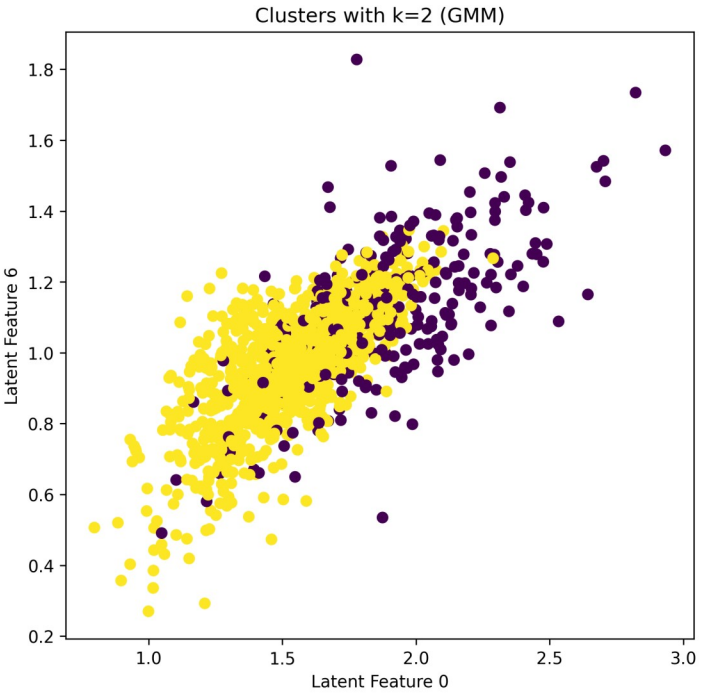
For example, for k=3, Cluster 0 is identified by having a high values for latent features, Cluster 1 having middling values and Cluster 2 having low values.

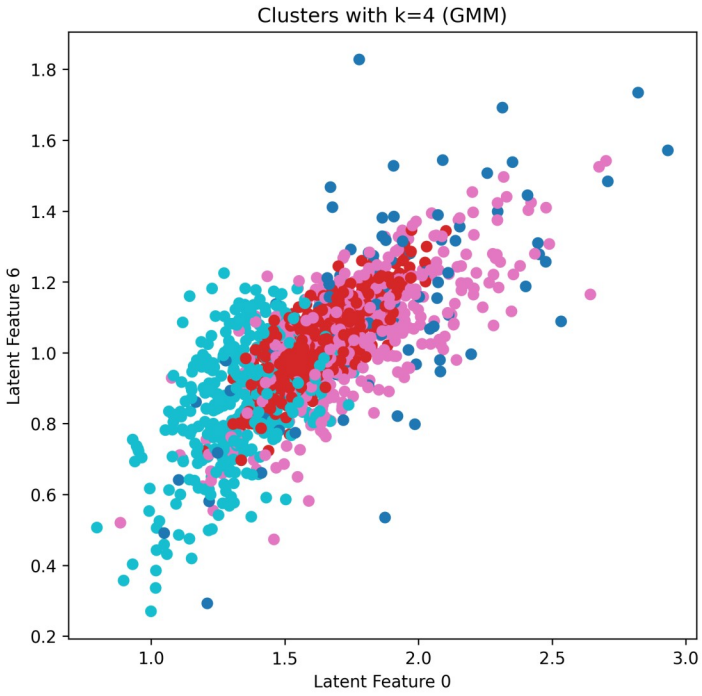
By contrast, for k=5 the clusters' characteristics are far less immediately obvious and rely on multiple variables to determine cluster

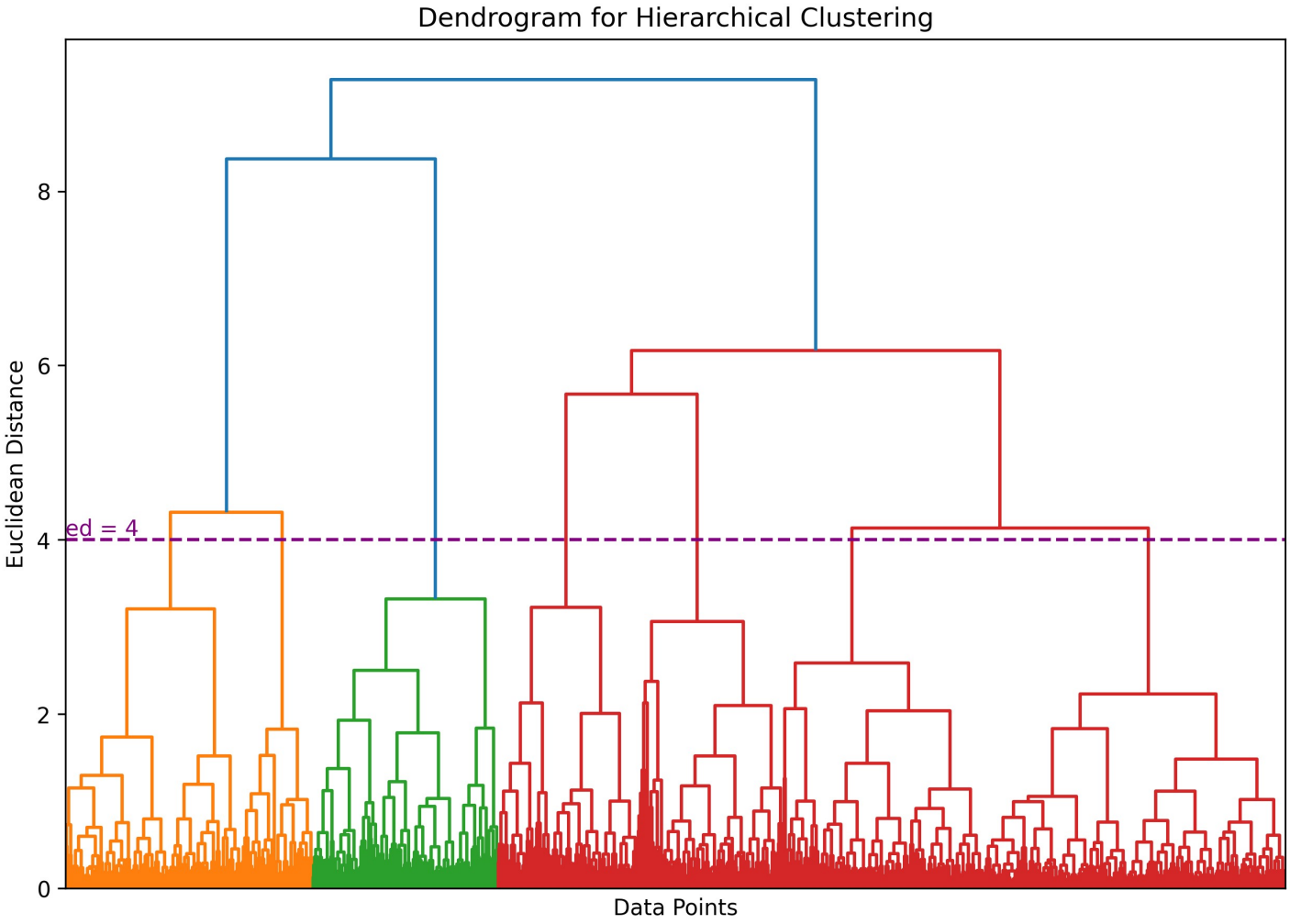




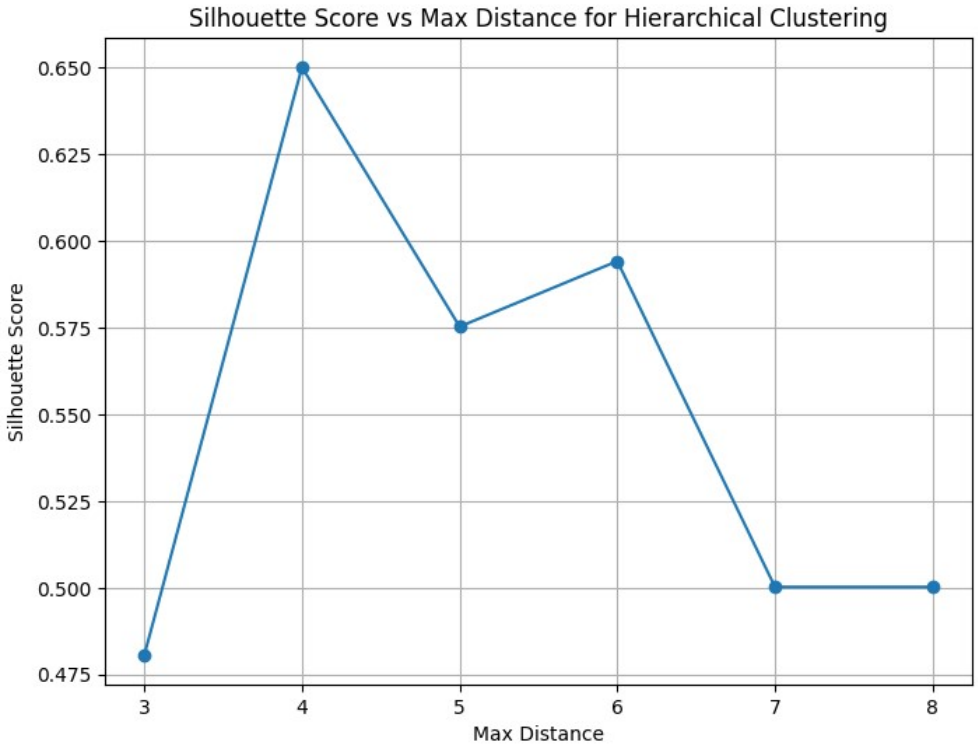
These graphs visualise the K-means clustering method (above) against the GMM clustering method (below). The data shows a 2D slice of the latent space, chosen to highlight the difference in clustering. As you can see the groups of points above are more well-defined and there is less overlap. This can be represented by the silhouette score, for which they are (0.30, 0.28) above and (0.19,0.09) below. For this reason, I chose to reject the clustering by GMM.

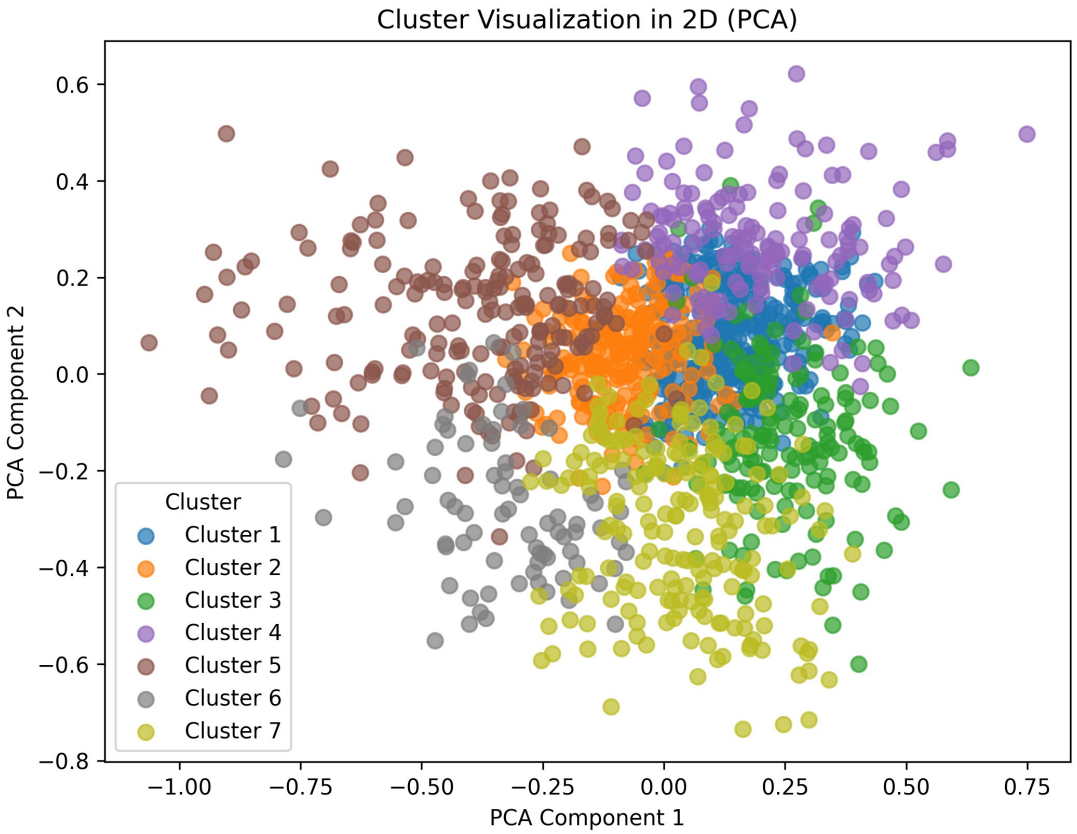




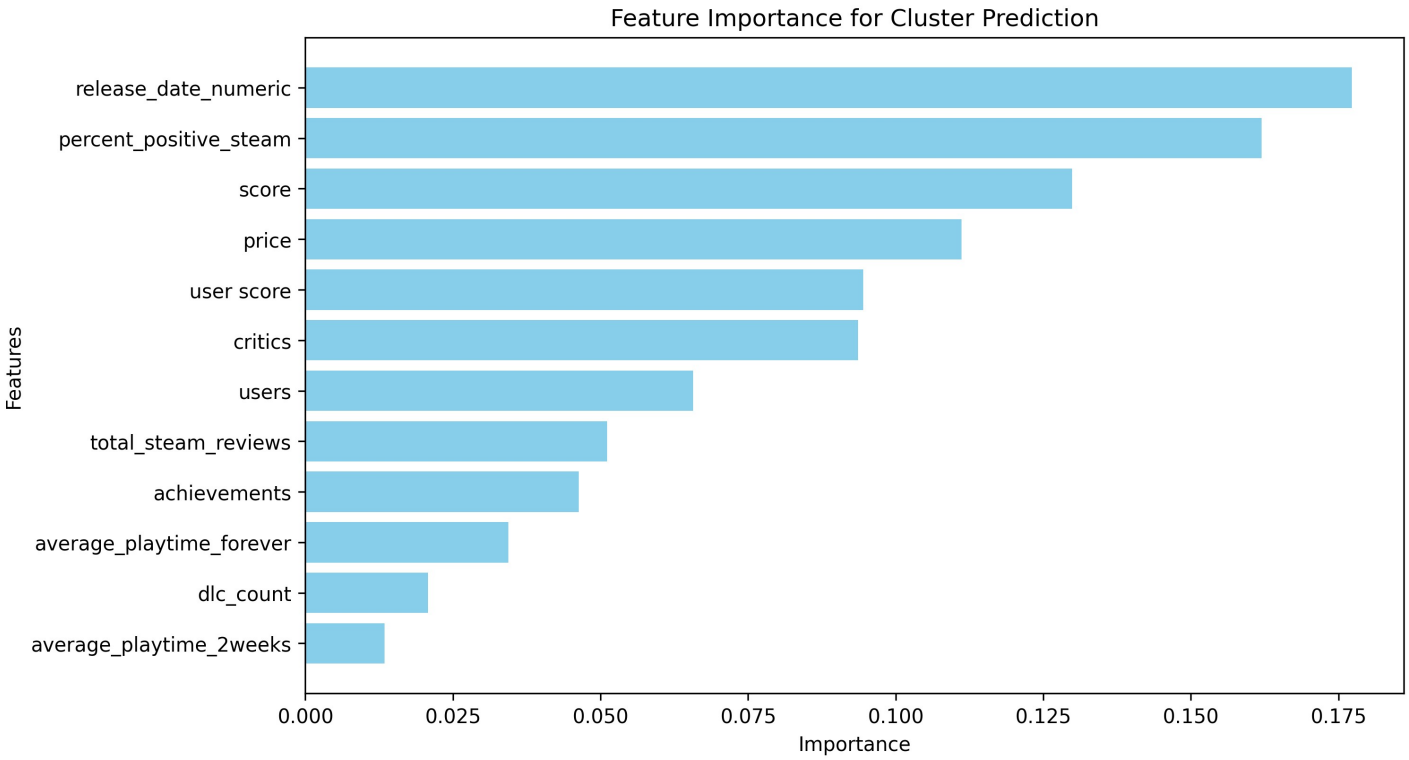


The above graph shows the dendrogram used in the process of hierarchical clustering. After computing the dendrogram, I computed the silhouette score as a function of distance, and used it to choose 4 as the distance. By drawing a horizontal line on the dendrogram at ed=4, we can see it makes 7 intersections corresponding to 7 clusters. Also note the high value of the silhouette score compared to previous methods, which is why I chose hierarchical clustering.

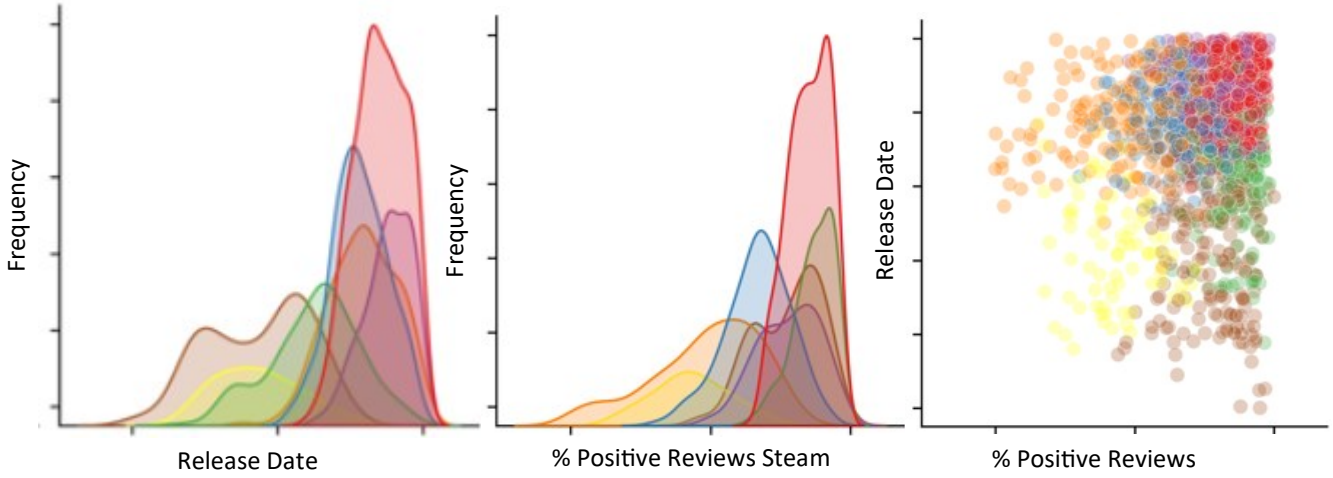




The above graph visualises the 7 clusters after doing PCA on the feature space to reduce it to 2 dimensions. As a result, it may seem that there is significant overlap, but that is largely due to the reduction of dimensionality of the space. We know this because of the high silhouette score from the clustering.



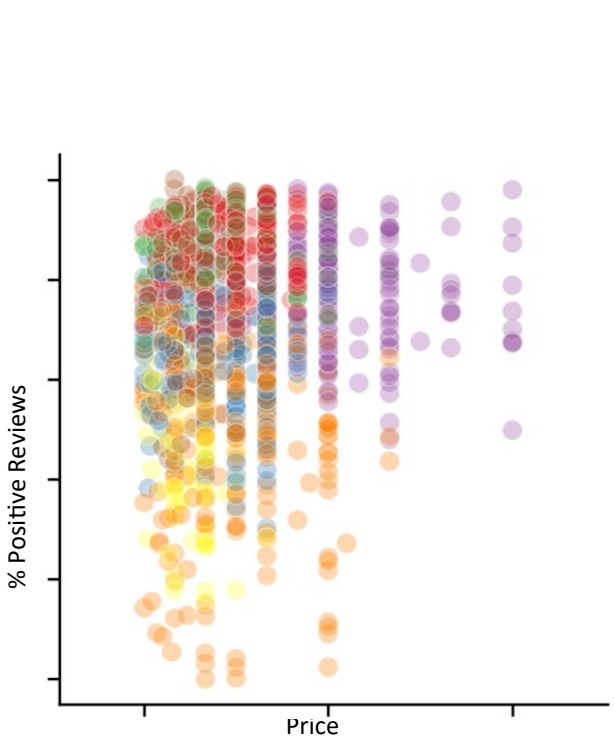
The above graph shows us the most important features for determining which cluster a data point belongs to. The most importance feature is given by release date, then by percentage positive reviews on steam and then by the Metacritic Critic Score. This graph is very useful when looking myopically at pair plots, as it can tell you which plots and relationships to look for.



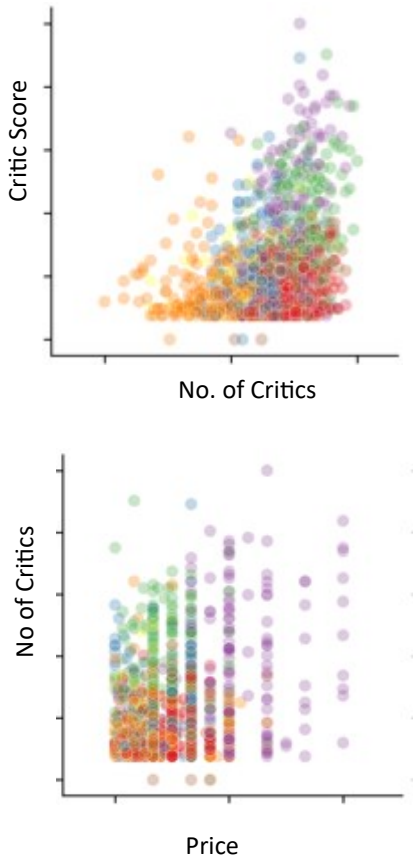
The left two graphs show the distribution for 'Release Date' and '% Positive Reviews Steam' for each cluster. For example, we can see games in the red, purple, blue and orange clusters have a more recent release date, whilst games in the yellow, brown and green clusters have older release dates. Similarly, we can see which clusters have a high percentage of positive reviews. The area coloured tells us the number of games within each cluster, e.g. the red cluster contains the most games.

However, looking at individual frequency diagrams often isn't enough to identify clusters. The plot on the right shows us the relationship between 'Release Date' and '% Positive Reviews Steam' by cluster. Now, we can start to see some separation.

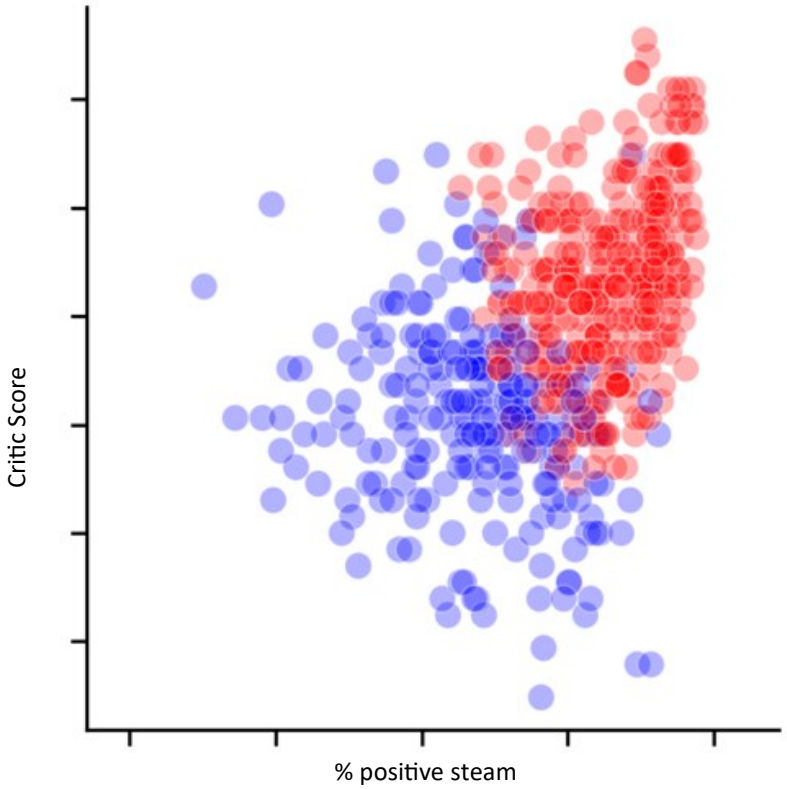
We can categorise that the yellow cluster on average contains games with an old release date and low percentage of positive reviews on Steam. Similarly, on average the orange cluster consists of games with a recent release date but low % of positive reviews. The brown cluster on average contains games with an old release date but a high % of positive reviews.



Looking at the above graph, we can categorise that on average games in the purple cluster will have a high price and a high % of positive reviews



We can see in the top graph that games in the green cluster have both a high critic score and number of critics alongside purple. However we can separate green and purple on the price as shown in the bottom graph.



Separating blue and red individually from the full pair plots were difficult to do visually but since we have categorised the other colours, we can plot only the red and blue clusters to see what distinguishes them from each other. From the above graph, so we can distinguish them from each other by looking at % positive reviews on steam and Critic Score from Meta-critic where red has high critic score and % positive reviews on steam, whereas blue may have at most a high score or high % positive reviews on steam but not both.

To summarise the identifiers of clusters, the table below:

Cluster Colour	Identifier
Yellow	Old Release Date + Low % Positive Reviews
Orange	Recent Release Date + Low % Positive Reviews
Brown	Old Release Date + High % Positive Reviews
Purple	High Price, High % Positive Reviews
Green	High Critic Number, High Critic Score, Low Price
Red	Separate from Blue by High Critic Score, High Positive % Reviews
Blue	Separate from Red

To conclude this section, after trying various clustering methods I chose hierarchical clustering for my categorisation, identified the important features used in making the categorisation and then identified the characteristics of each cluster.