A comparison of a dark vs light theme for a online book store

BS Steyn *(Department of Computer Science)*
Stellenbosch University
Stellenbosch, South Africa
21740178@sun.ac.za

I. Introduction

This report will focus on investigating the statistical significance of themes on a websites conversion rate. For the themes there are two option either light or dark. To most the option of which is better should be obvious as we want to keep those bugs away but we will look at this from a statistical view. The data we will use models a website wherein there exist the option to be in either light or dark mode. The dataset has the following features:

- Theme: dark or light.
- Click Through Rate: The proportion of the users who click on links or buttons on the website.
- Conversion Rate: The percentage of users who signed up on the platform after visiting for the first time.
- Bounce Rate: The percentage of users who leave the website without further interaction after visiting a single page.
- Scroll Depth: The depth to which users scroll through the website pages.
- Age: The age of the user.
- Location: The location of the user.
- Session Duration: The duration of the user's session on the website.
- Purchases: Yes or No
- Added to Cart: Yes or No

We will be focusing on the influence of 'Theme' on the 'Conversion Rate', the 'Click Through Rate' and on 'Purchases'. These options were chosen as these features are clearly connected to the engagement of the users and the financial effectiveness of the site. If it is found that there is no impact from the 'Theme' feature it would indicate that implementation of varying theme was not necessary and that the site should focus on other features that have impact. The second part of this report will focus on using k-means clustering to investigate if there is a pattern under the users. The results of this investigation will be interesting as it will show whether visuals and possibly first impressions(as it is not record whether this is a first visit or not) have a large influence on engagement and effectiveness. It will also be interesting to see whether there are definitive clusters in the data that patterns can be extracted from or whether no such pattern exists in the dataset.

II. Implementation

This project consists of two parts one focused on the statistical analysis of the dataset using multiple null hypotheses. The second focuses on using k-means clustering and PCA dimensionality reduction to create clusters and visualise them.

A. Part I

This part will focus on the implementation of the statistical analysis. The fundamental components of this part are as follows:

1) Descriptive statistical table
2) Null hypothesis for 'Theme' and 'Conversion Rate'
3) Null hypothesis for 'Theme' and 'Click Through Rate'
4) Null hypothesis for 'Theme' and 'Purchases'
5) Confidence interval for 'Conversion Rate'
6) Confidence interval for 'Click Through Rate'

The following subsections will provide an elaboration on these components listed above.

1) Descriptive statistical table: This table shows the mean, the standard deviation and variance for all features. The mean and standard deviation values were calculated using the python library Numpy. While the variance is the square of the standard deviation.

2) Null hypothesis for 'Theme' and 'Conversion Rate': The Null hypothesis is that the mean of the 'Conversion Rate' where the 'Theme' is dark is equal to the mean of the 'Conversion Rate' where the 'Theme' is light. This is tested using a Two-sample t-test. These are tested for 0.95 significance level. The normality and homogeneity are tested as well.

3) Null hypothesis for 'Theme' and 'Click Through Rate': The Null hypothesis is that the mean of the 'Conversion Rate' where the 'Theme' is dark is equal to the mean of the 'Click Through Rate' where the 'Theme' is light. This is tested using a Two-sample t-test. These are tested for 0.95 significance level. The normality and homogeneity are tested as well.

4) Null hypothesis for 'Theme' and 'Purchases': The Null hyphothesis is that 'Theme' and 'Purchases' are independent of each other. This is tested using a Chi-square test. These are tested for 0.95 significance level.

5) Confidence interval for 'Conversion Rate': The confidence interval for the mean of the 'Conversion Rate' is determined using scipy Python library. This is determined for 0.95 confidence value.

6) Confidence interval for 'Click Through Rate': The confidence interval for the mean of the 'Click Through Rate' is determined using scipy Python library. This is determined for 0.95 confidence value.

B. Part II

This part will focus on the clustering the data using k-means as well as using PCA to decrease the dimensionality of the data set for better visualisation. The fundamental components of this part are as follows:

1) The normalisation pipeline
2) The PCA dimensionality reduction
3) The k-means clustering
4) Evaluation pipeline
5) Visualisation

The following subsections will provide an elaboration on these components listed above.

1) The normalisation pipeline: The normalisation pipeline starts by one-hot encoding the 'Theme', 'Purchases' and 'Added to Cart' features. Then the 'Location' feature is binary encoded. After all categorical feature have been encoded the dataset is normalise so that they have a range of [-1,1]. This is done by dividing every observation by its maximum absolute value.

2) The PCA dimensionality reduction: After the normalisation pipeline PCA is used to decrease the dimensionality of the dataset to 3. This is done to allow for easier visualisation as well as allowing for better clustering.

3) The k-means clustering: The k-means clustering algorithm is used to create cluster labels for each entry in the dataset. To find the optimal value for k both the elbow method as well as the Calinski-Harabasz index are evaluated on k for a range of [2,30].

4) Evaluation pipeline: The final clusters then are assigned a silhouette score this is also how it was determine that clustering the reduced dimensionality dataset yielded better cluster compared to the non-reduced dataset.

5) Visualisation: This clusters found are visualised for for reduced dimensionality as well as non-reduced. The non-reduced are shown by plot a feature vs feature strip plot.

III. Results Part I

This section will look at the results of the implementation for Part I.

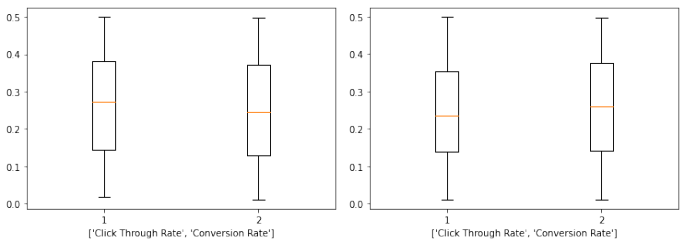
A. Descriptive statistical table

In the table below we will the descriptive statistic for all the non-categorical features on the dataset.

	Mean	Standard Deviation	Variance
Click Through Rate	0.25605	0.1391	0.019
Conversion Rate	0.25331	0.13902	0.019
Bounce Rate	0.50575	0.1721	0.0296
Scroll Depth	50.31949	16.8868	285.164
Age	41.528	14.1073	199.016
Session Duration	924.999	507.9775	258041.1405

TABLE I: Descriptive Statistical table

As shown in the table above due to the difference in magnitude of the dataset normalisation will be required in Part II. Boxplots for the 'Click Through Rate', 'Conversion Rate' for the different themes can be seen below:

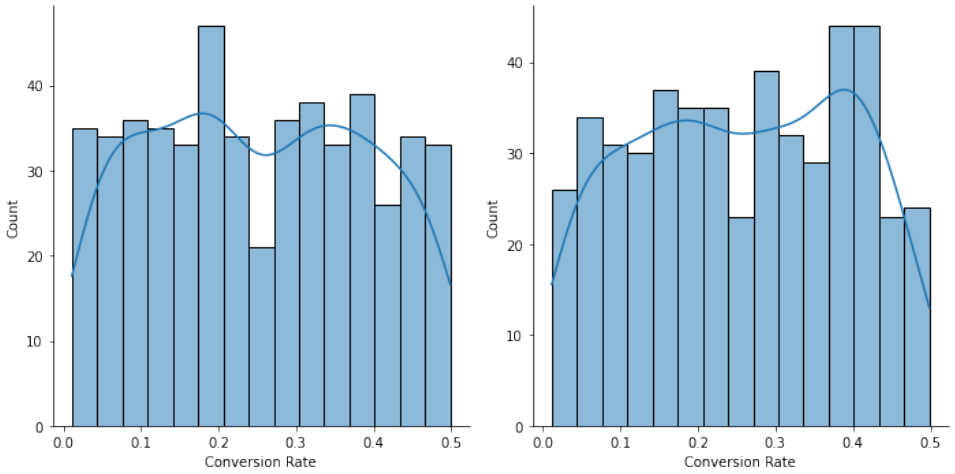


(a) Dark Boxplot (b) Light Boxplot

Fig. 1: Boxplots

B. Null hypothesis for 'Theme' and 'Conversion Rate'

For this section we start by testing the normality of the 'Conversion Rate' for dark and light. This is done by using another null hypothesis test using a t-test. Our null hypothesis for is that the data is normally distributed. Both light and dark reject the null hypothesis meaning that the data is not normally distribute however due the central limit theorem we can still preform our Two-sample t-test. The distribution can be seen below:



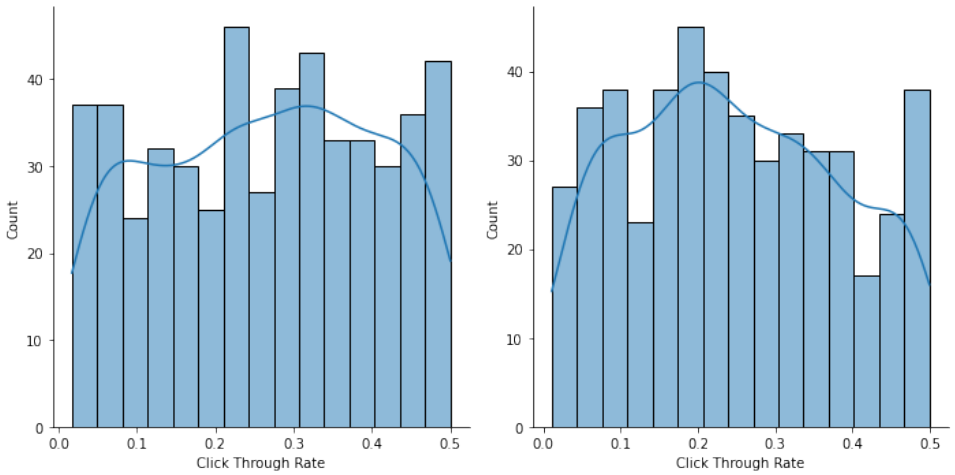
(a) Dark distribution (b) Light distribution

Fig. 2: Data Distribution for 'Conversion Rate'

Next we check the homogeneity of our data using another null hypothesis with a t-test. Our null hypothesis is that the variances of the light and dark 'Conversion Rate' data are equal. We fail to reject the null hypothesis meaning the variances of the samples are the same. Our p value was equal to 0.4208 Now finally move to our main question. With null hypothesis that the means of the samples are equal we preform our Two-sample t-test. It is found that we fail to reject the null hypothesis with a p value of 0.63525232. Meaning that 'Theme' does not have an impact on the 'Conversion Rate'.

C. Null hypothesis for 'Theme' and 'Click Through Rate'

For this section we start by testing the normality of the 'Click Through Rate' for dark and light. This is done by using another null hypothesis test using a t-test. Our null hypothesis for is that the data is normally distributed. Both light and dark reject the null hypothesis meaning that the data is not normally distribute however due the central limit theorem we can still preform our Two-sample t-test. The distribution can be seen below:



(a) Dark distribution (b) Light distribution

Fig. 3: Data Distribution for 'Click Through Rate'

Next we check the homogeneity of our data using another null hypothesis with a t-test. Our null hypothesis is that the variances of the light and dark 'Click Through Rate' data are equal. We fail to reject the null hypothesis meaning the variances

of the samples are the same. Our p value was equal to 0.3134 Now finally move to our main question. With null hypothesis that the means of the samples are equal we preform our Two-sample t-test. It is found that we reject the null hypothesis with a p value of 0.04835031. Meaning that 'Theme' does have an impact on the 'Click Through Rate'.

D. Null hypothesis for 'Theme' and 'Purchases'

Our null hypothesis is that 'Theme' and 'Purchases' are independent from each other hence meaning that 'Theme' does not impact 'Purchases'. We will use a Chi-squared test to test for independence. Our test Chi-test value is 0.6238119337882984 with a p value of 0.4296342647705996 and degrees of freedom 1. This means that we fail to reject the null hypothesis based on the p value. Meaning that 'Purchases' are independent of 'Theme'.

E. Confidence interval for 'Conversion Rate'

The confidence interval for 'Conversion Rate' where the 'Theme' is Dark is (0.2390683823024078, 0.2634955650180487) while the calculated mean is 0.25128. The confidence interval for 'Conversion Rate' where the 'Theme' is Light is (0.24322944415817188, 0.26768918997785646) while the calculated mean is 0.2555. This was done for a confidence value of 0.95.

F. Confidence interval for 'Click Through Rate'

The confidence interval for 'Click Through Rate' where the 'Theme' is Dark is (0.252296245799803, 0.27670547912990323) while the calculated mean is 0.2645. The confidence interval for 'Click Through Rate' where the 'Theme' is Light is (0.23488419828318088, 0.2593332233704358) while the calculated mean is 0.2471. This was done for a confidence value of 0.95.

IV. Result Part II

This section will look at the results of the implementation for Part II.

A. The normalisation pipeline

The results of the normalisation can be see in the figure below. The figure will show the first 5 entries before and after normalisation.

	Theme	Click Through Rate	Conversion Rate	Bounce Rate	Scroll_Depth	Age	Location	Session_Duration	Purchases	Added_to_Cart
0	1	0.054920	0.282367	0.405085	72.489458	25	Chennai	1535	0	1
1	1	0.113932	0.032973	0.732759	61.858568	19	Pune	303	0	1
2	0	0.323352	0.178763	0.296543	45.737376	47	Chennai	563	1	1
3	1	0.485836	0.325225	0.245001	76.305298	58	Pune	385	1	0
4	1	0.034783	0.196766	0.765100	48.927407	25	New Delhi	1437	0	0

Fig. 4: Original Data Table

	Theme	Click Through Rate	Conversion Rate	Bounce Rate	Scroll_Depth	Age	Session_Duration	Purchases	Added_to_Cart	Bangalore	Chennai	Kolkata	New Delhi	Pune
0	1.0	0.109842	0.565961	0.506573	0.906151	0.384615	0.854201	0.0	1.0	0.0	1.0	0.0	0.0	0.0
1	1.0	0.227869	0.066090	0.916341	0.773260	0.292308	0.168614	0.0	1.0	0.0	0.0	0.0	0.0	1.0
2	0.0	0.646718	0.358304	0.370837	0.571738	0.723077	0.313300	1.0	1.0	0.0	1.0	0.0	0.0	0.0
3	1.0	0.971694	0.651863	0.306383	0.953851	0.892308	0.214246	1.0	0.0	0.0	0.0	0.0	0.0	1.0
4	1.0	0.069569	0.394387	0.956785	0.611615	0.384615	0.799666	0.0	0.0	0.0	0.0	0.0	1.0	0.0

Fig. 5: Normalised Data Table

The effect of the encoding and normalisation can be seen in the figures above.

B. The PCA dimensionality reduction

In this section the dataset is transformed using PCA to reduce its dimensionality. A visualisation of the data in it's new form can be seen in the figure below:

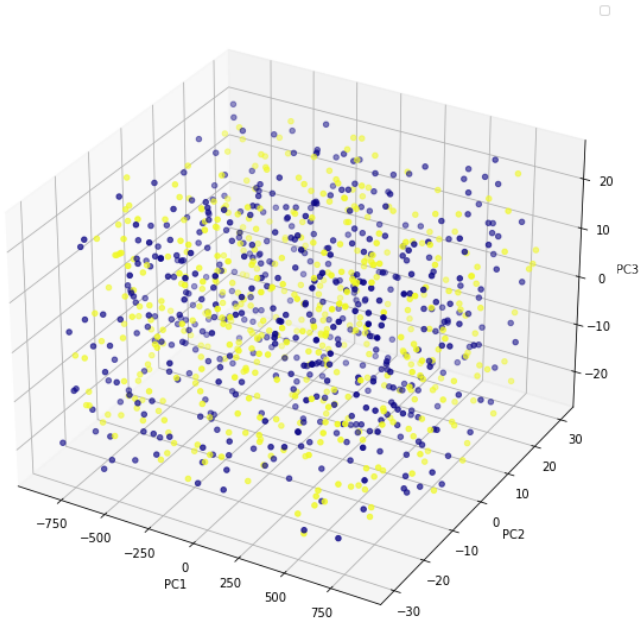
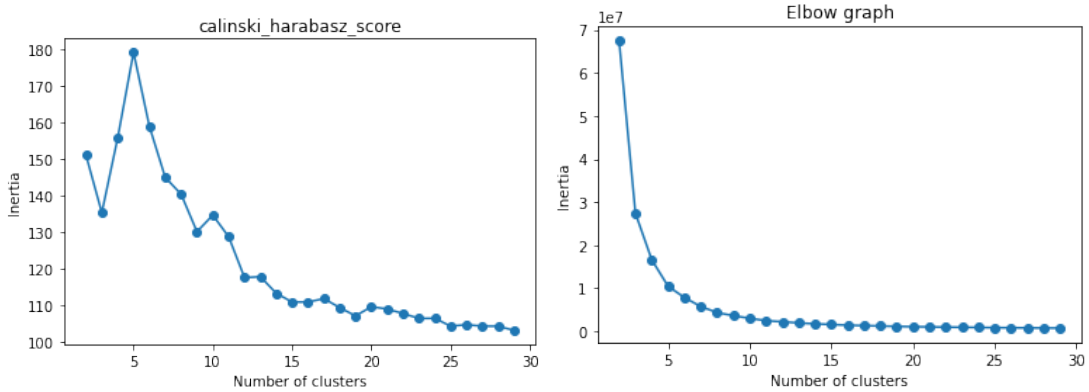


Fig. 6: PCA dataset

Where the yellow dots represent entries where the 'Theme' is Light and the blue dots represent the entries where 'Theme' is Dark.

C. The k-means clustering

In this section we will show the results optimised k-means clustering. To start we begin by determining what the optimal value is for k. This is done using the elbow method and the Calinski and Harabasz score. The results for a range of [2,30) can be seen in the figure below:



(a) Calinski (b) Elbow

Fig. 7: Optimal k tests

The figure above shows us that both the elbow method and Calinski and Harabasz score agree that k=5 is the optimal k value. This is the value that will be used for the next section. Visualisation of the clusters will be shown in the final subsection.

D. Evaluation pipeline

This section shows for all test k values as well as the optimal k value. The values can be seen in the figure below:

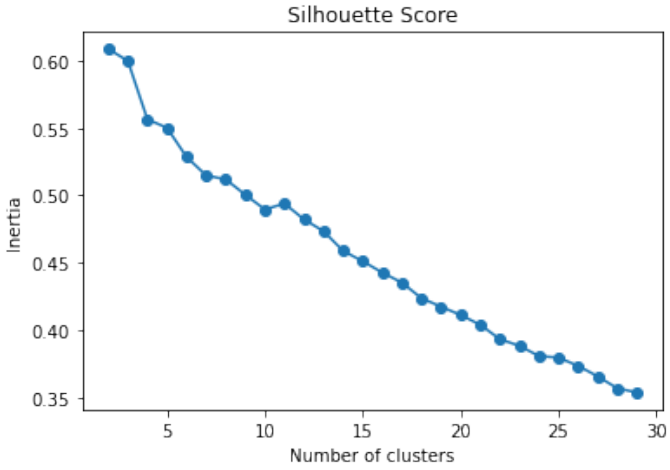


Fig. 8: Silhouette Scores

The silhouette score obtained for our optimal k is 0.552. This is massive improvement over the silhouette score that we obtain without applying PCA which was around 0.2. This shows that the clusters created using the reduced dimensionality dataset are far better.

E. Visualisation

This section is comprised of various figures showing the created clusters.

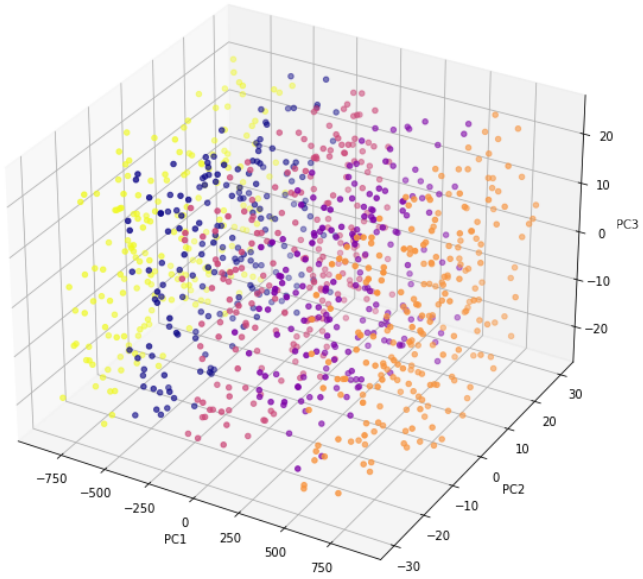


Fig. 9: Cluster Visualisation

We can see that there are clear clusters created in the reduce dimensionality set however we would like to see how these clusters look in our original dataset this can be done by viewing a strip plot of all the features which can be seen below.

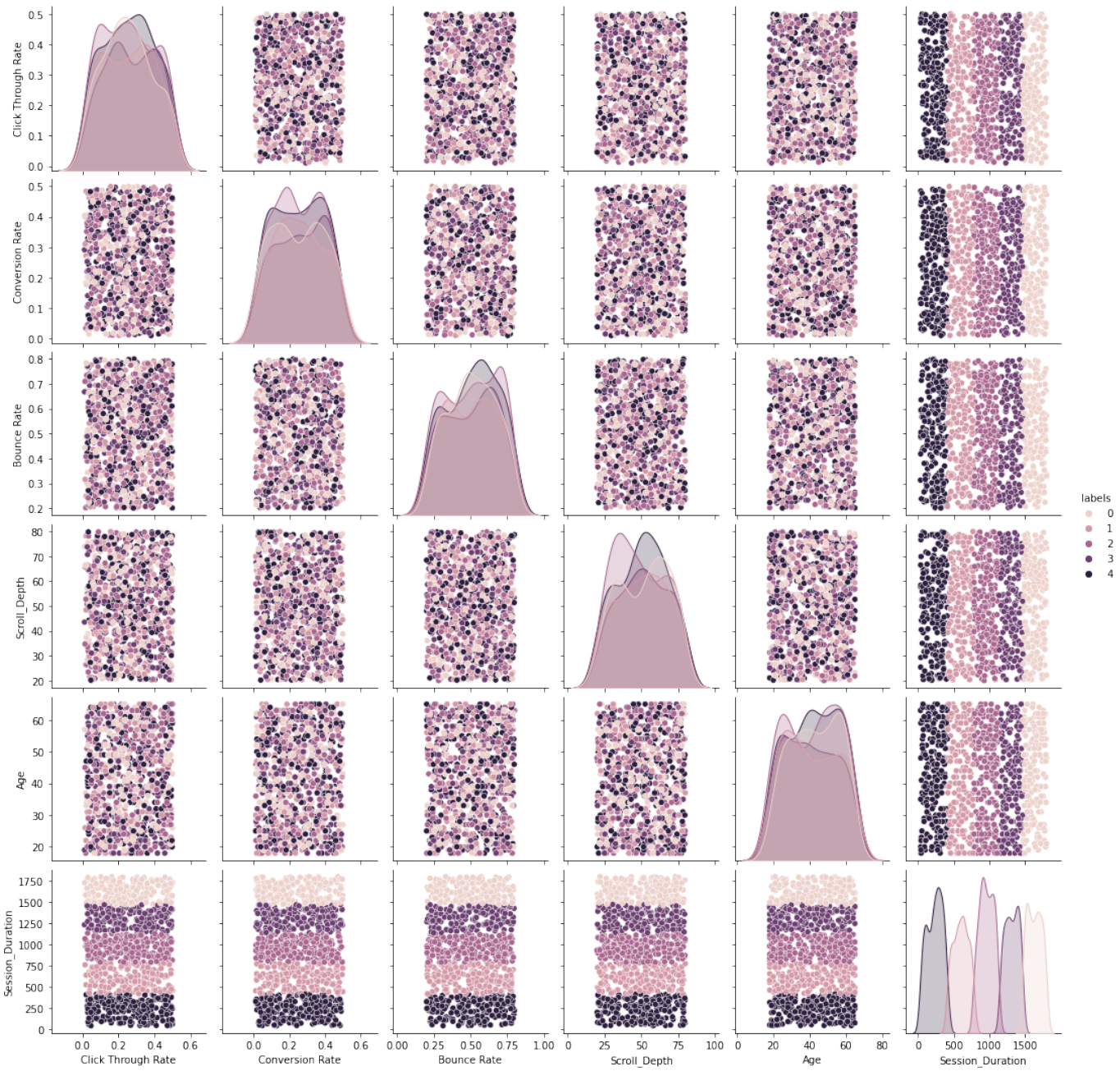


Fig. 10: Cluster Visualisation

From this figure it is clear to see that the feature with the largest impact on the clusters in 'Session Duration'.

V. Conclusion

For Part I 'Theme' does not effect all the users to a great degree as only 'Click Through Rate' was effect from the features investigated. Part II revealed a interesting connection between 'Session Duration' and the clusters in the dataset. Further investigation into 'Session Duration' relationships could be an interesting next step.