UNIVERSITY OF ECONOMICS AND LAW

COURSE: BIG DATA ANALYTICS

CLUSTERING E-COMMERCE CUSTOMERS USING K-MEANS ALGORITHM

GROUP 4

List of members:

K194141725 Thái Tuấn Kha (Team leader)

K194141751 Lâm Nhựt Thịnh

K194141723 Nguyễn Tuấn Hưng

K194141755 Trần Thanh Trúc

Supervisor:

Nguyen Thon Da Ph.D

Ho Chi Minh City, 2022

TABLE OF CONTENTS

ABSTRACT

This study deploys the topic of clustering e-commerce customers. The problem is performed on a dataset of customers' e-commerce activities. This dataset collected from the Harvard Dataverse website consists of about 500 observations. After processing data and applying K-means algorithm, the customers in the dataset were divided into 5 separate clusters based on their “Time on Website”, “Time on App”, “Length of Membership” and “Yearly Amount Spent”.

1. Introduction

1.1 Research topic

This research paper solves the problem of clustering e-commerce customers based on their interaction behavior on the platform. The main idea of this research paper is formed from the common customer segmentation problem of businesses in the context of the rapidly growing global e-commerce market.

1.2 Dataset

The dataset used in this research is collected from Harvard Dataverse website. This dataset contains 500 observations and 7 columns. Its columns can be explained as follow:

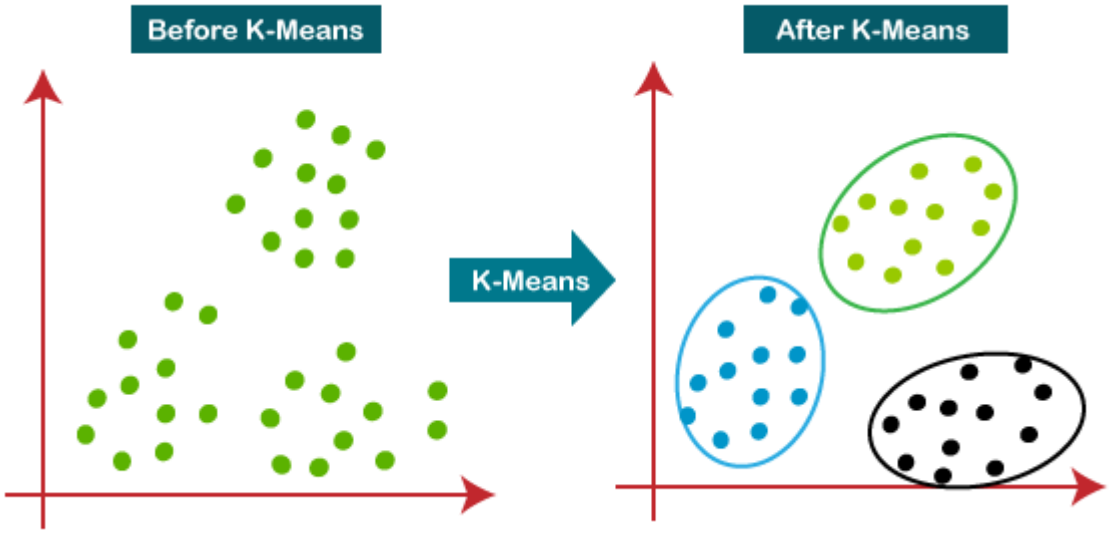
- Email: Personal email of customers
- Address: Personal address of customers
- Avatar: The main color of customer account's avatar on the platform
- Time on Website: The amount of time (hours) the customer spent on the website of the platform in one year.
- Time on App: The amount of time (hours) the customer spent on the app of the platform in one year.
- Length of Membership: Length of the time they joined the e-commerce platform.
- Yearly Amount Spent: The amount of money (USD) the customer spent on buying products from the platform in one year.

1.3 Tools and algorithms used

We use Google Colab as the coding platform, data processing and analysis as well as algorithm implementation will be performed on it.

The problem in this study is solved by the K-means algorithm. K-Means Clustering is an Unsupervised Learning algorithm, which groups the unlabeled dataset into different clusters. Here K defines the number of pre-defined clusters that need to be created in the process, as if K=2, there will be

two clusters, and for K=3, there will be three clusters, and so on. It allows us to cluster the data into different groups and a convenient way to discover the categories of groups in the unlabeled dataset on its own without the need for any training.



2. Relate work

Since ecommerces are growing a lot in recent years, there are many previous studies built the Machine Learning models to predict the yearly amount spent by customers.

The study led by Guilherme Abreu (2021) builds a ML model that uses linear regression to predict the yearly amount spent by customers in an ecommerce. He finds that for each unity growth on “Time on App”, the “Yearly Amount Spent” increase in 38 Dollars. In addition, the membership is what causes the most increase, but, maybe, focus on the App could give a good result too, maybe it’s cheaper to improve the app rather than increase the membership length.

The study led by Hannah Sophia Seippel (2018) analyzes machine learning models to predict a purchase, which is a relevant use case as applied by a large German clothing retailer. Next, to comparing models this study further gives insight into the performance differences of the models on sequential clickstream and the static customer data, by conducting a descriptive data

analysis and separately training the models on the different datasets. The results indicate that a Random Forest algorithm is best suited for the prediction task, showing the best performance results, reasonable latency, offering comprehensibility and a high robustness.

Another study led by Thomas Wood (2019) uses a predictive modelling through machine learning. He used Python and AI to predict how much customers will spend in a store the next time they visit. He added more input features to the Regression model such as “day of week”, “day of year”, “isChristmasSeason” etc. Moreover, he switched to a Polynomial Regression Model and Random Forest Regression. This allowed him to find the optimal model for his study. Finally, he finds that Random Forest Regressor is the best model and its mean absolute error reaches to 20.5.

3. Preliminaries

Customer segmentation is a common problem of many businesses. Segmentation of customer groups can help managers allocate strategies more effectively to each customer group. E-commerce is a growing field in the world with a huge number of customers. Therefore, the problem of customer segmentation is even more urgent for businesses in this field. This study will simulate and implement the idea of customer clustering based on the data of their e-commerce activity.

4. Data Exploration and Data Visualization

In this part, we will look into performing EDA on the data using PySpark

4.1 Data Acquisition

PySpark has its own dataframe and functionalities. Here we red a single CSV into the dataframe using `spark.read.csv()` and then use that dataframe to perform the analysis. We set the **inferSchema** attribute as True, this will go through the CSV file and automatically adapt its schema into PySpark Dataframe. We get the following output for the above steps:

```
+--------------------+--------------------+---------------+-----------+---------------+------------------+-----------------
-+
|               Email|             Address|         Avatar|Time on App|Time on Website|Length of Membership|Yearly Amount Spen
t|
+--------------------+--------------------+---------------+-----------+---------------+------------------+-----------------
-+
|mstephenson@ferna...|835 Frank TunnelW...|         Violet|      12.66|          39.58|              4.08|            587.9
5|
|   hduke@hotmail.com|4547 Archer Commo...|       DarkGreen|      11.11|          37.27|              2.66|             392.
2|
|    pallen@yahoo.com|24645 Valerie Uni...|         Bisque|      11.33|          37.11|               4.1|            487.5
5|
|riverarebecca@gma...|1414 David Throug...|     SaddleBrown|      13.72|          36.72|              3.12|            581.8
5|
|mstephens@davidso...|14023 Rodriguez P...|MediumAquaMarine|       12.8|          37.54|              4.45|            599.4
1|
|alvareznancy@luca...|645 Martha Park A...|     FloralWhite|      12.03|          34.48|              5.49|             637.
1|
|katherine20@yahoo...|68388 Reyes Light...|   DarkSlateBlue|      11.37|          36.68|              4.69|            521.5
7|
|  awatkins@yahoo.com|Unit 6538 Box 898...|            Aqua|      12.35|          37.37|              4.43|             549.
9|
|vchurch@walter-ma...|860 Lee KeyWest D...|         Salmon|      13.39|          37.53|              3.27|             570.
2|
|    bonnie69@lin.biz|PSC 2734, Box 525...|          Brown|      11.81|          37.15|               3.2|             427.
2|
+--------------------+--------------------+---------------+-----------+---------------+------------------+-----------------
-+
only showing top 10 rows
```

4.2 Data Description

df.describe().show() allows us to see a general description of the dataset, similar to what pandas dataframe.describe() does. We get the following output:

```
+-------+-----------------+--------------------+-----------+------------------+------------------+--------------------+--------
-----------+
|summary|            Email|             Address|     Avatar|       Time on App|   Time on Website|Length of Membership|Yearly A
mount Spent|
+-------+-----------------+--------------------+-----------+------------------+------------------+--------------------+--------
-----------+
|  count|              500|                 500|        500|               500|               500|                 500|
500|
|   mean|             null|                null|       null|12.052619999999996| 37.06048000000002|  3.5333599999999983| 499.314
24000000004|
| stddev|             null|                null|       null|0.9944179552579887|1.0105553987921965|  0.9992604030512829|  79.314
76398498823|
|    min|aaron04@yahoo.com|0001 Mack MillNor...|  AliceBlue|              8.51|             33.91|                0.27|
256.67|
|    max|zscott@wright.com|Unit 7502 Box 834...|YellowGreen|             15.13|             40.01|                6.92|
765.52|
+-------+-----------------+--------------------+-----------+------------------+------------------+--------------------+--------
-----------+
```

We further use df.printSchema() to get the schema information of the dataset:

4.3 Data Preprocessing

Pyspark dataframes don't have readily available, easy to run functions to check for data inconsistencies and null values. We are going to use driver function, with the help of Pyspark's inbuilt SQL functions to check for the null values in our dataframe:

```
# Creating a dataframe to check null value counts
null_df = df.select([count(when(col(c).contains('None') | \
                            col(c).contains('NULL') | \
                            (col(c) == '') | \
                            col(c).isNull() | \
                            isnan(c), c
                           )).alias(c)
                     for c in df.columns])

# Displaying the null value counts dataframe
null_df.show()
```

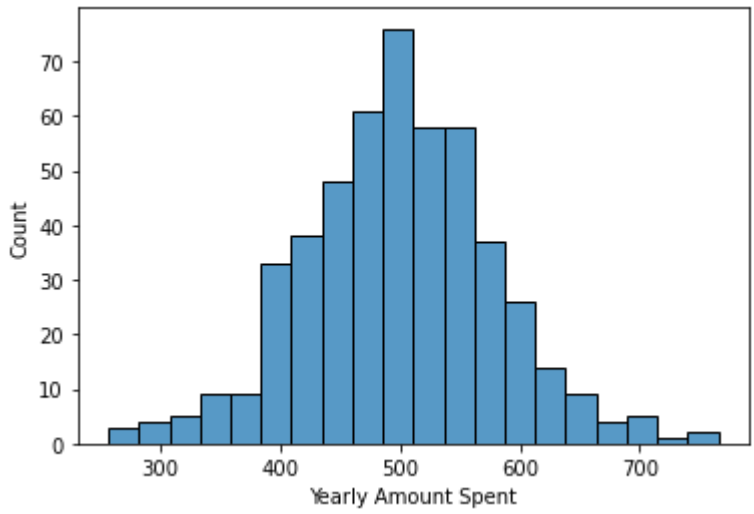
```
+-----+-------+------+----------+---------------+-------------------+------------------+
|Email|Address|Avatar|Time on App|Time on Website|Length of Membership|Yearly Amount Spent|
+-----+-------+------+----------+---------------+-------------------+------------------+
|    0|      0|     0|         0|              0|                  0|                 0|
+-----+-------+------+----------+---------------+-------------------+------------------+
```

Clearly, the dataset doesn't have null values, so it doesn't need any preprocessing and we can simply proceed with the EDA of this dataset.

4.4 Exploratory Data Analysis

PySpark dataframes do not support visualizations like pandas does with its plot() method. We will simply convert our pyspark dataframe to **pandas** dataframe. Then, we will import one advanced plotting library – Seaborn that provides a histplot() function that can be used to plot histograms. Some of the univariate analysis that uses this histplot() function method is as follows:

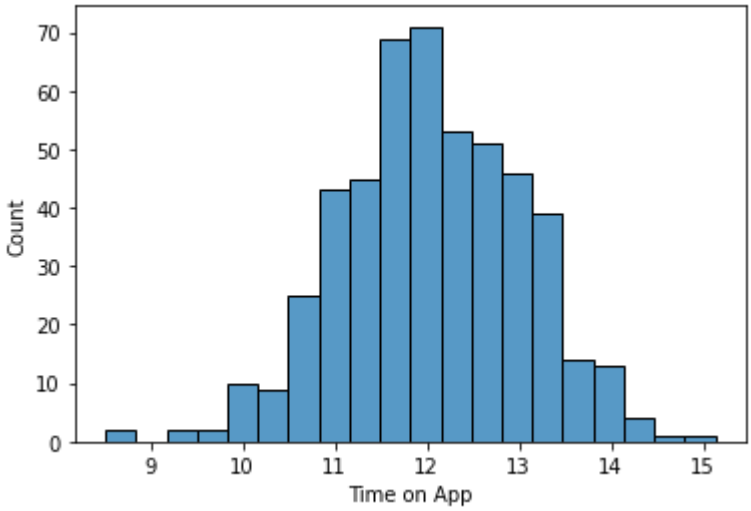
Question 1: What is the spread of the *Yearly Amount Spent* feature?



- We can see it closely looks like a **symmetrical distribution**.

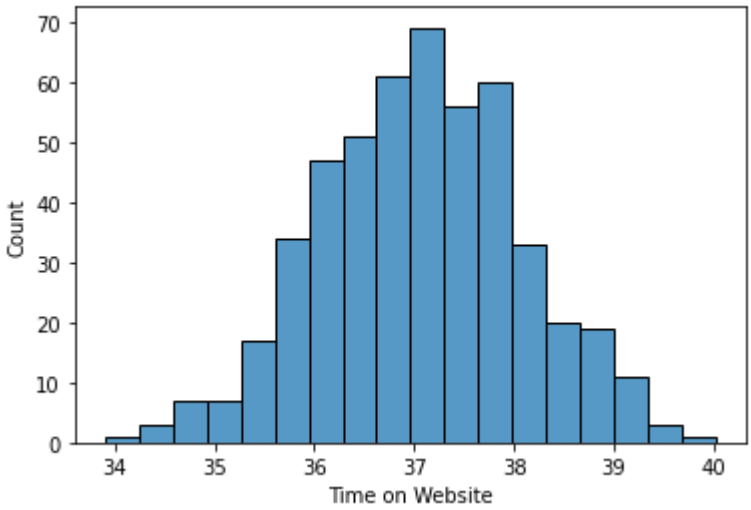
- **Max Yearly Amount Spent** can go till **765 dollars**.
- **The majority** of the customer have Yearly Amount Spent listings **above 400 dollars** and **under 600 dollars**.

Question 2: What is the spread of the *Time on App* feature?



- We can see it closely looks like a **symmetrical distribution**.
- **Max Time on App** can go till 16 hours.
- **The majority** of the customers have Time on App listings **above 11 hours** and **under 14 hours**.

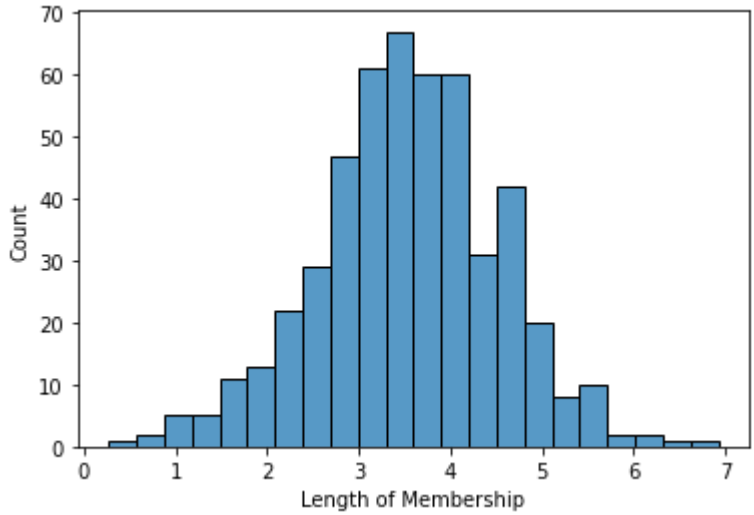
Question 3: What is the spread of the *Time on Website* feature?



- We can see it closely looks like a **symmetrical distribution**.
- **Max Time on Website** can go till 16 hours.

- **The majority** of the customer have Time on Website listings **above 35 hours** and **under 39 hours**.

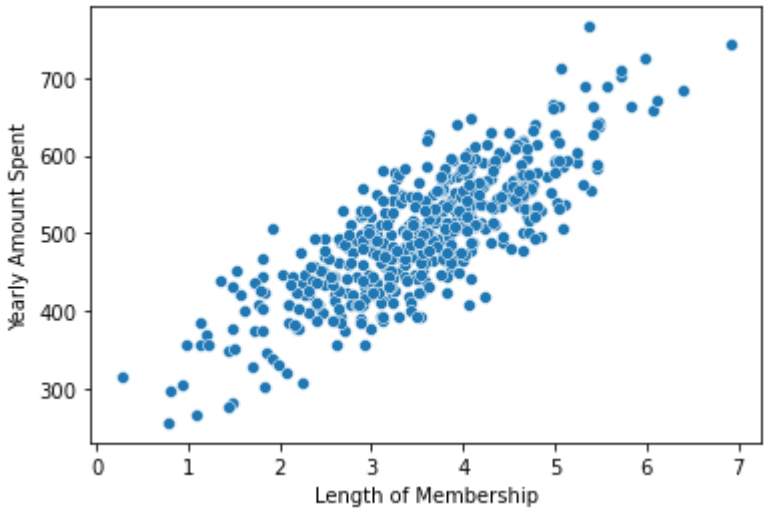
Question 3: What is the spread of the *Length of Membership* feature?



- We can see it closely looks like a **symmetrical distribution**.
- **Max Length of Membership** can go till 16 hours.
- **The majority** of the customers have Length of Membership **above 2 months** and **under 5 months**.

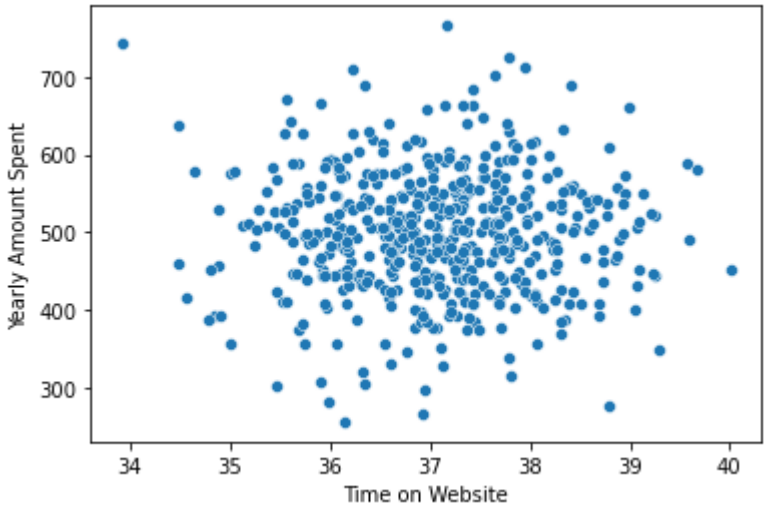
Let's see some Multivariate Analysis. Seaborn allows you to plot a scatterplot with its scatterplot() function. Let's see some examples:

Question 4: Determine the relation between *Yearly Amount Spen*t and *Length of Membership* features?



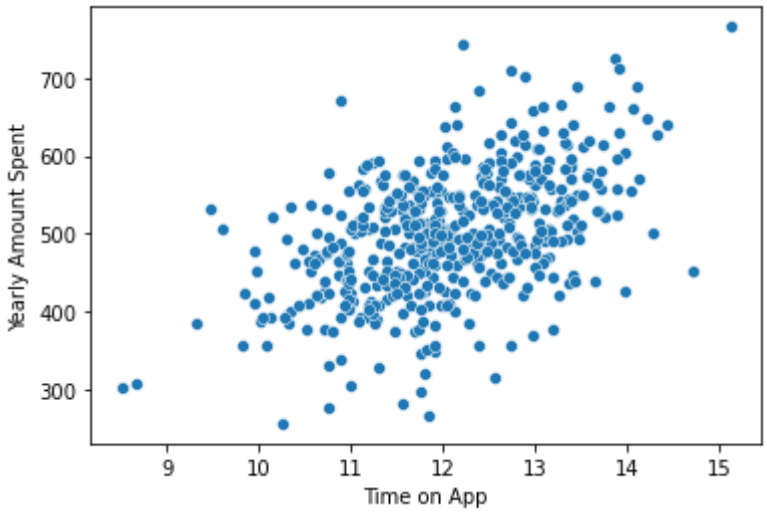
We observe a strong linear relationship between **Yearly Amount Spent** and **Length of Membership**.

Question 5: Determine the relation between *Yearly Amount Spent* and *Time on Website* features?



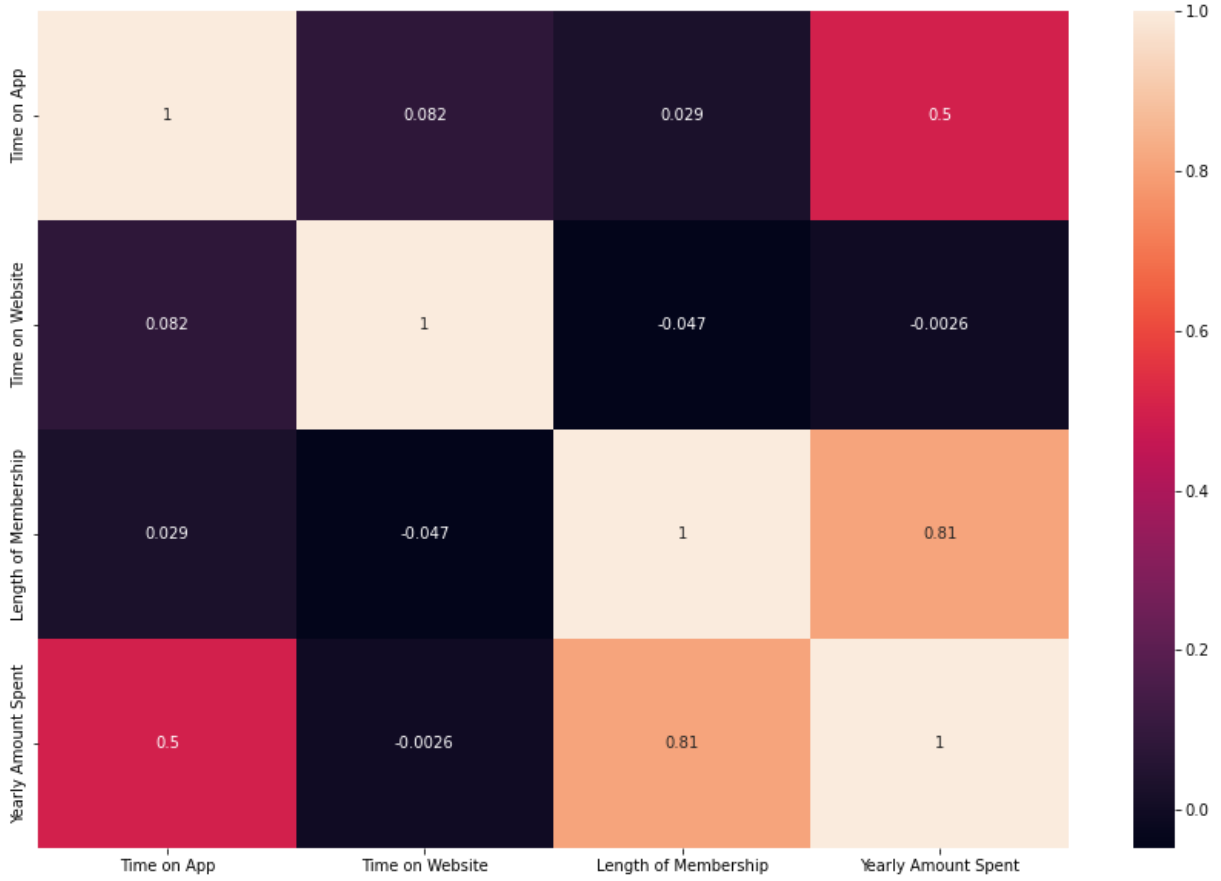
We don't observe a strong linear relationship between **Time on Website** and **Yearly Amount Spent**.

Question 6: Determine the relation between *Yearly Amount Spent* and *Time on App* features?



We don't observe a strong linear relationship between **Time on App** and **Yearly** Amount Spent.

Question 7: Plot a heatmap to check for correlations between features?



- **Yearly Amount Spent** is highly correlated with **Length of Membership** and **Time on App**.
- We can see **Yearly Amount Spent** is slightly related to **Time on Website** but is not influenced as much.
- **Time on Website** and **Length of Membership** are inversely correlated, indicating that some customers who have a longer length of membership usually spend less time on the website.

4.5 Conclusion

This Dataset contains 8 columns and 500 rows. After doing some analysis, I found out that Length of membership is the feature that have the biggest impact at the yearly amount spent by customers. After that, I started ploting some graphs to see this relation better. The plot showed a positive relation between length of membership and yearly amount spent. This means that the longer the membership, the highest the amount spent.

5. Proposed model

K-Means clustering model will be applied to solve the clustering problem mentioned in this study. The reason for choosing this model is because the problem to be solved is defined as an unsupervised learning problem. K-Means clustering is the most basic algorithm of unsupervised learning, approaching the problem with an uncomplicated algorithm will avoid getting highly complex results in the research process.

6. Experimental results

The number of clusters (k) is the most important indicator of the K-means model. Determining the right value of k helps optimize the process of clustering. In this study, Silhouette score will be used to find the optimal value of k.

Silhouette Score = (b-a)/max(a,b), where:

a= average intra-cluster distance i.e the average distance between each point within a cluster.

b= average inter-cluster distance i.e the average distance between all clusters.

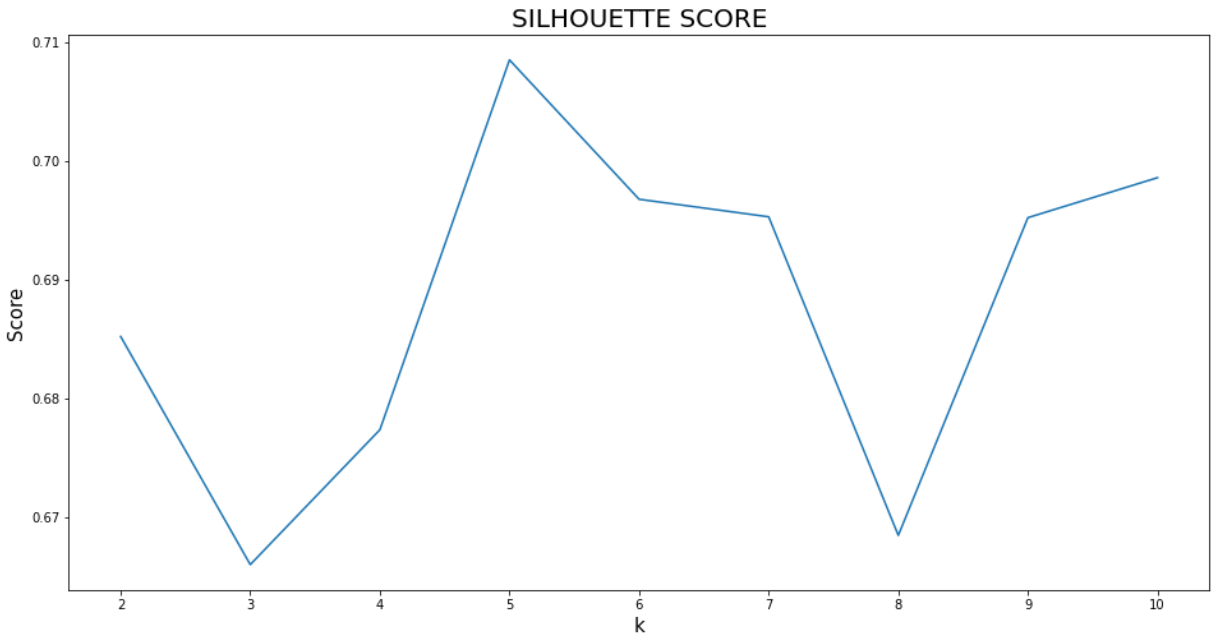
Silhouette score is a metric used to calculate the goodness of a clustering technique. Its value ranges from -1 to 1.

1: Means clusters are well apart from each other and clearly distinguished.

0: Means clusters are indifferent, or we can say that the distance between clusters is not significant.

-1: Means clusters are assigned in the wrong way.

The chart giving information about Silhouette score corresponding with the value of k is shown below.

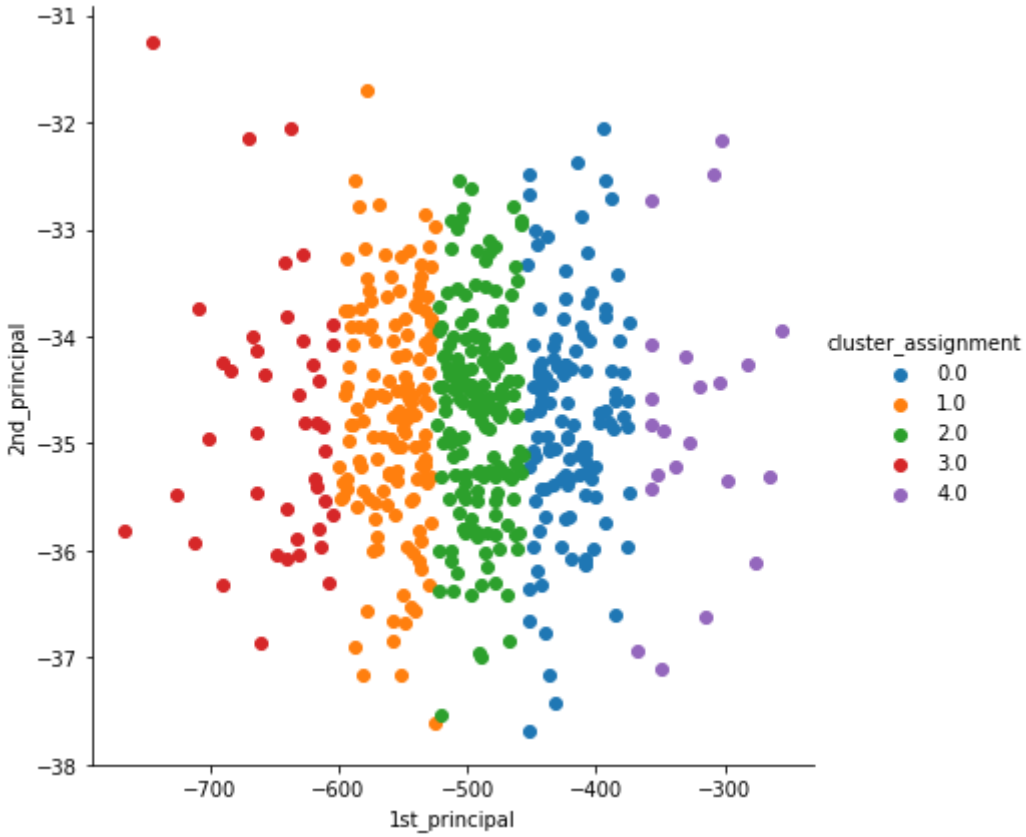


It can be seen from the above chart that the Silhouette score peaked when k = 5, so the optimal value of k (number of clusters) is 5.

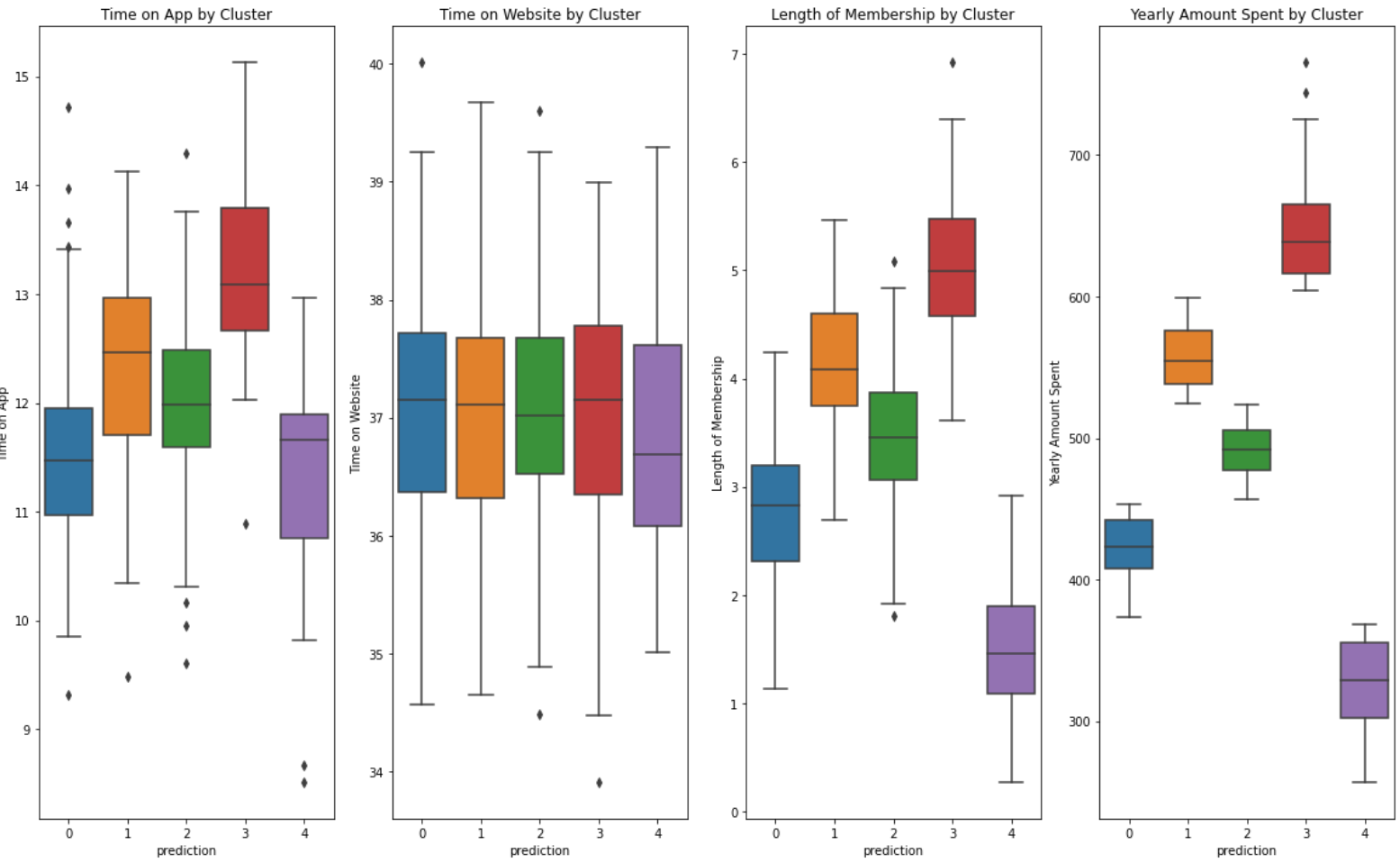
After determining the appropriate value of k (k=5), the K-means model is deployed and this is the cluster centers:

```
Center point of cluster 1:  [ 11.52239669  37.1092562    2.75826446 420.8653719 ]
Center point of cluster 2:  [ 12.36703448  37.07413793   4.11868966 557.63448276]
Center point of cluster 3:  [ 11.99523529  37.05952941   3.48047059 491.08811765]
Center point of cluster 4:  [ 13.15785714  36.97928571   5.01452381 649.00595238]
Center point of cluster 5:  [ 11.23        36.86454545   1.51954545 324.19      ]
```

Then visualize the clustering result in detail for a closer look.



It can obviously be seen that the customers are divided into 5 clusters represented by 5 different colors. Clusters are quite close to each other but they do not overlap. It is important to examine carefully the characteristics of each cluster. The below plots can help us to do that.



There are some insights drawn from the above plots:

- There is not much difference in Time on Website between clusters.
- The red cluster contains the most loyal customers of this e-commerce platform. They have the highest Yearly Amount Spent, the highest Length of Membership and they also spend the most time on the app.
- The purple cluster contains the least interactive customers of this e-commerce platform. They have the lowest Yearly Amount Spent, the lowest Length of Membership and they also spend the least time on the app.
- The value of features of the remaining 3 clusters lies between that of the red cluster and the purple cluster. The value of features of these 3 clusters is not too different from each other, the orange cluster is slightly higher than the green cluster and the blue cluster.

Based on the above insights, there are some advice that the ecommerce platform can follow:

- Loyal customer group (like red cluster or orange cluster) is a very important customer group for businesses. They need to have methods to retain this group of customers.
- For a group of customers who interact poorly (like purple cluster or blue cluster), businesses need to attract this audience by incentives, promotions or gifts.
- The website of this e-commerce platform has not been effective in stimulating the shopping behavior of customers, while the app of this platform has a good effect because the customers spending more time on the app tend to buy more.

7. Conclusion

In the scope of this research, the customers have been divided into 5 fairly clear clusters by K-means algorithm. However, the fact that most variables are highly correlated leads to the failure to uncover special insights. Limited data is the most likely cause for this problem. However, the idea of this research paper is completely reasonable as well as practical for e-commerce businesses.