**Customer-Segmentation-Using-Spark**

The primary aim of this project is to apply advanced data analytics using PySpark techniques to gain a deeper understanding of mall customers and their shopping behaviors.

**Key Insights**

Customer Segmentation: The clustering process categorized mall customers into five distinct segments. These segments provide a deeper understanding of customer behaviors and preferences based on gender, age, annual income, and spending score.

Cluster Insights

**Cluster 1**: Customers in this group have relatively low annual incomes but are willing to spend, indicating potential for targeted promotions and loyalty programs.

**Cluster 2**: High-income customers with very high spending scores represent valuable shoppers who may respond well to premium offerings.

**Cluster 3**: Customers in this cluster exhibit moderate income and moderate spending patterns, making them a balanced target for marketing strategies.

**Cluster 4**: High-income shoppers with very high spending scores, suggesting opportunities for upscale products and personalized experiences.

**Cluster 5**: Customers in this cluster have lower income and spending scores, indicating more conservative spending habits.

**Silhouette Score**: The calculated Silhouette score of approximately 0.6276 indicates that the clusters are well-defined and provide meaningful insights.

**Actionable Insights**

With these customer segments, the mall can tailor marketing campaigns, product offerings, and promotions to better meet the needs and preferences of each cluster. This personalization can lead to improved customer engagement and increased sales.

**Report on Challenges Faced:**

1.\*Data Preprocessing Challenges\*:

- Missing or incomplete data: Often datasets contain missing values, which may need imputation or removal.

- Data inconsistencies: Some fields might have inconsistent formats or errors that need to be fixed.

2. \*Data Size and Performance\*:

- Handling large datasets efficiently using PySpark (it’s designed for big data, but you might face performance issues depending on the size of the data).

- Memory management in Spark, especially while working with large DataFrames.

3. \*Integration with Pandas\*:

- Visualizing data with PySpark: PySpark does not have robust built-in plotting functions, so converting data to Pandas for visualization can sometimes result in performance issues if the dataset is very large.

4. \*Spark Cluster Setup\*:

- Setting up Spark properly on your local machine or on a cluster might be tricky if you're not familiar with Spark or Hadoop ecosystems.

5. \*Complexity of PySpark APIs\*:

- PySpark’s API can sometimes be a bit challenging to get used to, especially when transitioning from Pandas or SQL.

6. \*Insight Extraction\*:

- Deriving meaningful insights from the data and finding the most relevant trends from the spending score.

**Analysis report on the basis of the spending score**

***Group by gender and find*** ***average spending score.***

+------+------------------+

| Genre| avg(Score)|

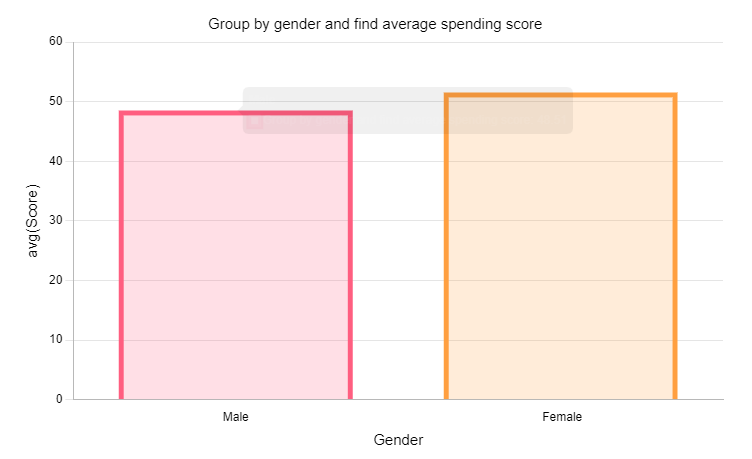
+------+------------------+

|Female|51.526785714285715|

| Male| 48.51136363636363|

+------+------------------+

Female customers have higher average spending score as compared to male customers.



***Filter out customers having spending score greater than 50.***

+----------+------+---+------------------+-----+

|CustomerID| Genre|Age|Annual Income (k$)|Score|

+----------+------+---+------------------+-----+

| 2| Male| 21| 15| 81|

| 4|Female| 23| 16| 77|

| 6|Female| 22| 17| 76|

| 8|Female| 23| 18| 94|

| 10|Female| 30| 19| 72|

| 12|Female| 35| 19| 99|

| 14|Female| 24| 20| 77|

| 16| Male| 22| 20| 79|

| 18| Male| 20| 21| 66|

| 20|Female| 35| 23| 98|

| 22| Male| 25| 24| 73|

| 24| Male| 31| 25| 73|

| 26| Male| 29| 28| 82|

| 28| Male| 35| 28| 61|

| 30|Female| 23| 29| 87|

| 32|Female| 21| 30| 73|

| 34| Male| 18| 33| 92|

| 36|Female| 21| 33| 81|

| 38|Female| 30| 34| 73|

| 40|Female| 20| 37| 75|

| 42| Male| 24| 38| 92|

| 44|Female| 31| 39| 61|

| 46|Female| 24| 39| 65|

| 47|Female| 50| 40| 55|

| 51|Female| 49| 42| 52|

| 52| Male| 33| 42| 60|

| 53|Female| 31| 43| 54|

| 54| Male| 59| 43| 60|

| 59|Female| 27| 46| 51|

| 61| Male| 70| 46| 56|

| 62| Male| 19| 46| 55|

| 63|Female| 67| 47| 52|

| 64|Female| 54| 47| 59|

| 65| Male| 63| 48| 51|

| 66| Male| 18| 48| 59|

| 69| Male| 19| 48| 59|

| 71| Male| 70| 49| 55|

| 74|Female| 60| 50| 56|

| 76| Male| 26| 54| 54|

| 77|Female| 45| 54| 53|

| 79|Female| 23| 54| 52|

| 81| Male| 57| 54| 51|

| 82| Male| 38| 54| 55|

| 85|Female| 21| 54| 57|

| 87|Female| 55| 57| 58|

| 88|Female| 22| 57| 55|

| 89|Female| 34| 58| 60|

| 91|Female| 68| 59| 55|

| 96| Male| 24| 60| 52|

| 103| Male| 67| 62| 59|

+----------+------+---+------------------+-----+

only showing top 50 rows

The above table shows customers having spending score greater than 50.

***Average spending score for people having age less than 35.***

+------------------+

| avg(Score)|

+------------------+

|60.449438202247194|

+------------------+

The average spending score for people having age less than is 60.449

***Average annual income for customers having spending score greater than 50.***

+-----------------------+

|avg(Annual Income (k$))|

+-----------------------+

| 60.41237113402062|

+-----------------------+

Average annual income for customers having spending score greater than 50 is 60k dollars.

***Top*** ***10 customers with highest spending score.***

+----------+------+---+------------------+-----+

|CustomerID| Genre|Age|Annual Income (k$)|Score|

+----------+------+---+------------------+-----+

| 12|Female| 35| 19| 99|

| 20|Female| 35| 23| 98|

| 146| Male| 28| 77| 97|

| 186| Male| 30| 99| 97|

| 128| Male| 40| 71| 95|

| 168|Female| 33| 86| 95|

| 8|Female| 23| 18| 94|

| 142| Male| 32| 75| 93|

| 164|Female| 31| 81| 93|

| 34| Male| 18| 33| 92|

+----------+------+---+------------------+-----+

only showing top 10 rows

The above table shows top 10 customers with highest spending score.

***Top 10 customers with lowest spending score.***

+----------+------+---+------------------+-----+

|CustomerID| Genre|Age|Annual Income (k$)|Score|

+----------+------+---+------------------+-----+

| 157| Male| 37| 78| 1|

| 159| Male| 34| 78| 1|

| 9| Male| 64| 19| 3|

| 31| Male| 60| 30| 4|

| 33| Male| 53| 33| 4|

| 135| Male| 20| 73| 5|

| 163| Male| 19| 81| 5|

| 141|Female| 57| 75| 5|

| 23|Female| 46| 25| 5|

| 3|Female| 20| 16| 6|

+----------+------+---+------------------+-----+

only showing top 10 rows

The above table shows 10 customers with lowest spending score.

***Group by age and find average spending score.***

+---+------------------+

|Age| avg(Score)|

+---+------------------+

| 31| 63.875|

| 65| 43.5|

| 53| 25.0|

| 34| 39.2|

| 28| 70.0|

| 27|60.333333333333336|

| 26| 54.5|

| 44| 13.5|

| 22| 70.0|

| 47| 28.5|

| 52| 21.0|

| 20| 40.2|

| 40| 47.5|

| 57| 28.0|

| 54| 35.75|

| 48| 41.8|

| 19| 39.75|

| 64| 3.0|

| 41| 28.0|

| 43| 34.0|

+---+------------------+

only showing top 20 rows

The above table shows customers grouped by age with their average spending score.

***Group by age class and average spending score for each class.***

+--------+------------------+

|AgeGroup| avg(Score)|

+--------+------------------+

| 49-58|36.964285714285715|

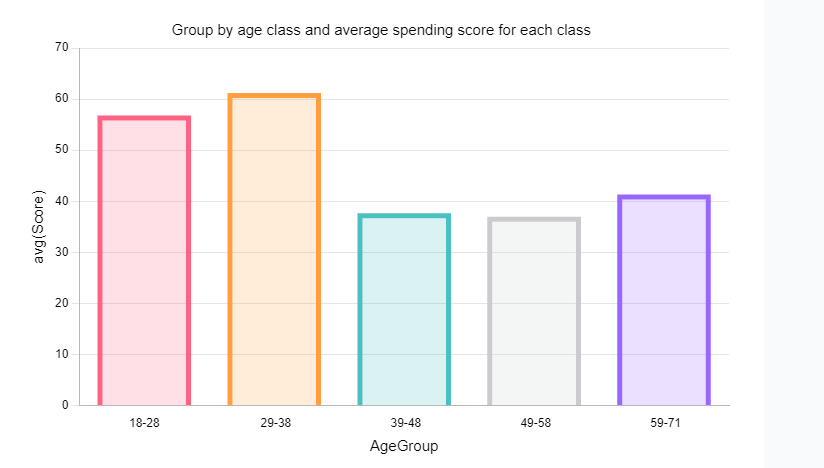
| 39-48| 37.65714285714286|

| 59-71|41.333333333333336|

| 18-28| 56.78|

| 29-38| 61.20634920634921|

+--------+------------------+



***Customer with maximum spending score.***

+----------+------+---+------------------+-----+

|CustomerID| Genre|Age|Annual Income (k$)|Score|

+----------+------+---+------------------+-----+

| 12|Female| 35| 19| 99|

+----------+------+---+------------------+-----+

Customer with customer ID 12 has the max. spending score.

***Customer with minimum spending score.***

+----------+-----+---+------------------+-----+

|CustomerID|Genre|Age|Annual Income (k$)|Score|

+----------+-----+---+------------------+-----+

| 157| Male| 37| 78| 1|

| 159| Male| 34| 78| 1|

+----------+-----+---+------------------+-----+

Customers with customer ID 157 and 159 have the min. spending score.

***Key Takeaways:***

1. Age group of 29-38 years has the maximum average spending score.
2. Female customers have higher average spending score as compared to male customers.
3. Customer with ID 12 has the maximum spending score.
4. Customers with ID 157 and 159 have the minimum spending score.
5. Average annual income for customers having spending score greater than 50 is 60k dollars.