CLUSTERING – MALL CUSTOMER SEGMENTATION

PUNEETHKRISHN B

DSFT8

INTRODUCTION:

- CUSTOMERS WHO COME FOR SHOPPING IN A MALL CAN BE GROUPED IN A NUMBER OF WAYS LIKE SHOPPING PATTERN, SPENDING PATTERN, SHOPPING SCORE, SALARY ETC.
- THE REASON FOR THESE CLUSTERS IS TO HELP IDENTIFY THOSE CUSTOMERS WHO WOULD BE INTERESTED IN CERTAIN PRODUCTS, OFFERS AND SERVICES.
- THE STORE MAY STRATEGIZE ITS OFFERINGS IN SUCH A WAY THAT IT TARGETS ONLY RIGHT CUSTOMERS FOR SPECIFIC PRODUCTS.

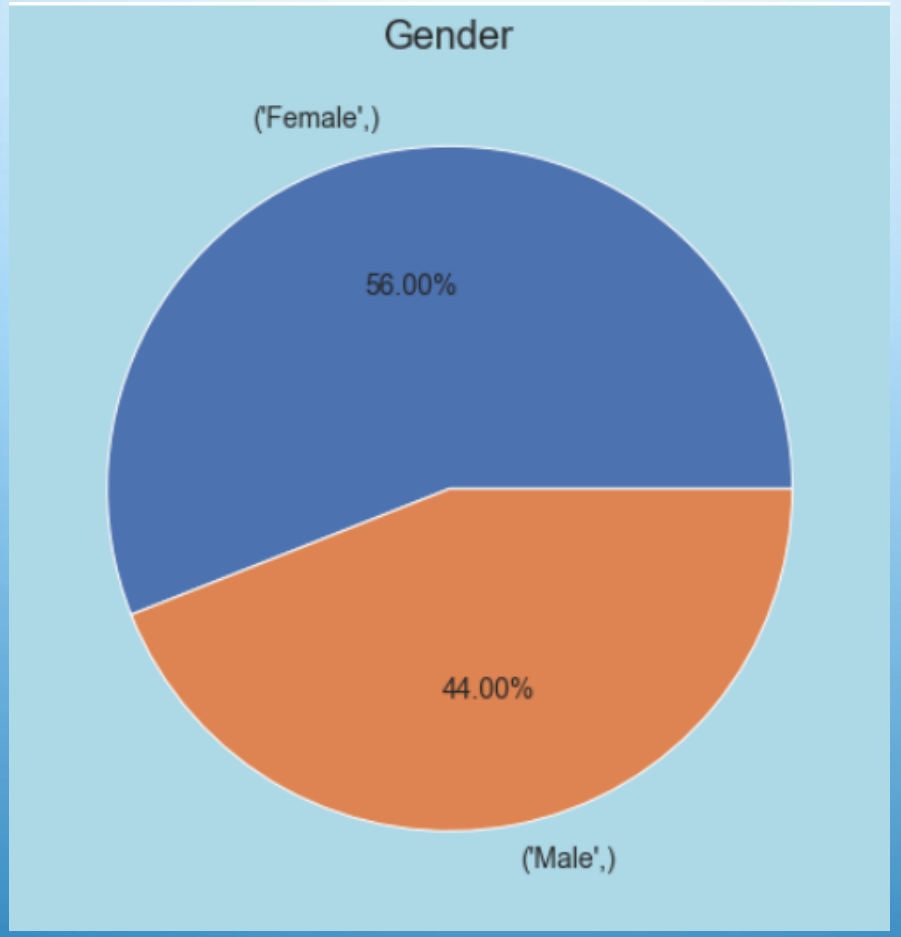
OBJECTIVE:

- TO HELP THE STORES BY CREATING CUSTOMER CLUSTERS TO IDENTIFY THE DIFFERENT TYPES OF CUSTOMERS

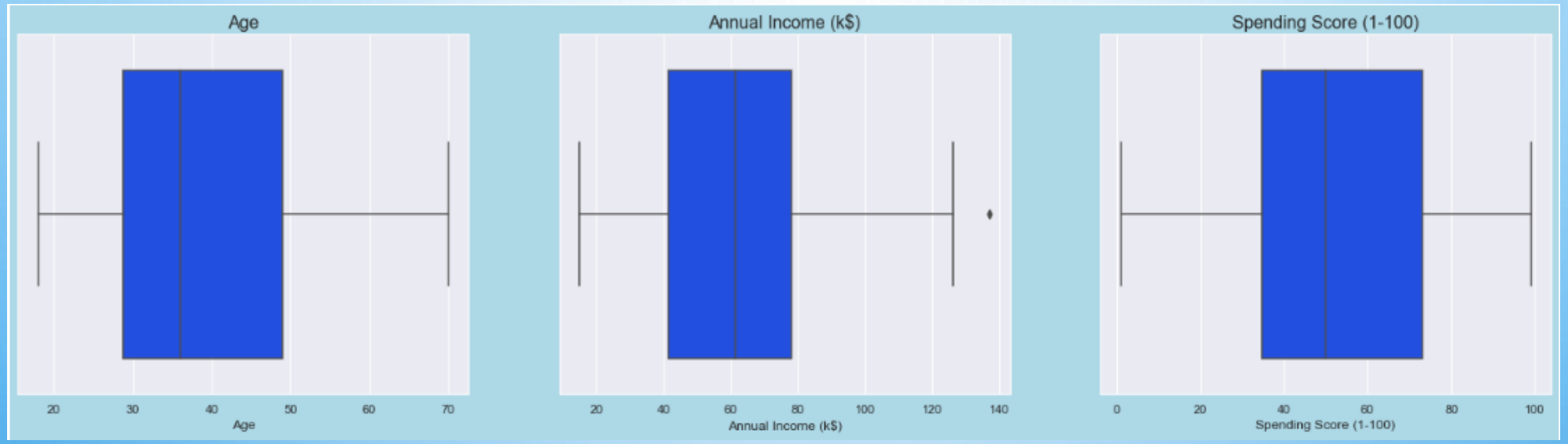
FEATURE	DATA TYPE	DESCRIPTION
Customer ID	Numeric	Unique customer ID
Gender	Character	Gender of the customer
Age	Numeric	Age of the customer
Annual income	Numeric	Annual income of the customer
Spending score	Numeric	Spending scores of customer to purchase products

➢DATA PRE-PROCESSING

- THERE IS NO ANY MISSING VALUES OR ANY SPECIAL CHARACTERS PRESENT IN THE DATASET
- **UNIVARIANT ANALYSIS**

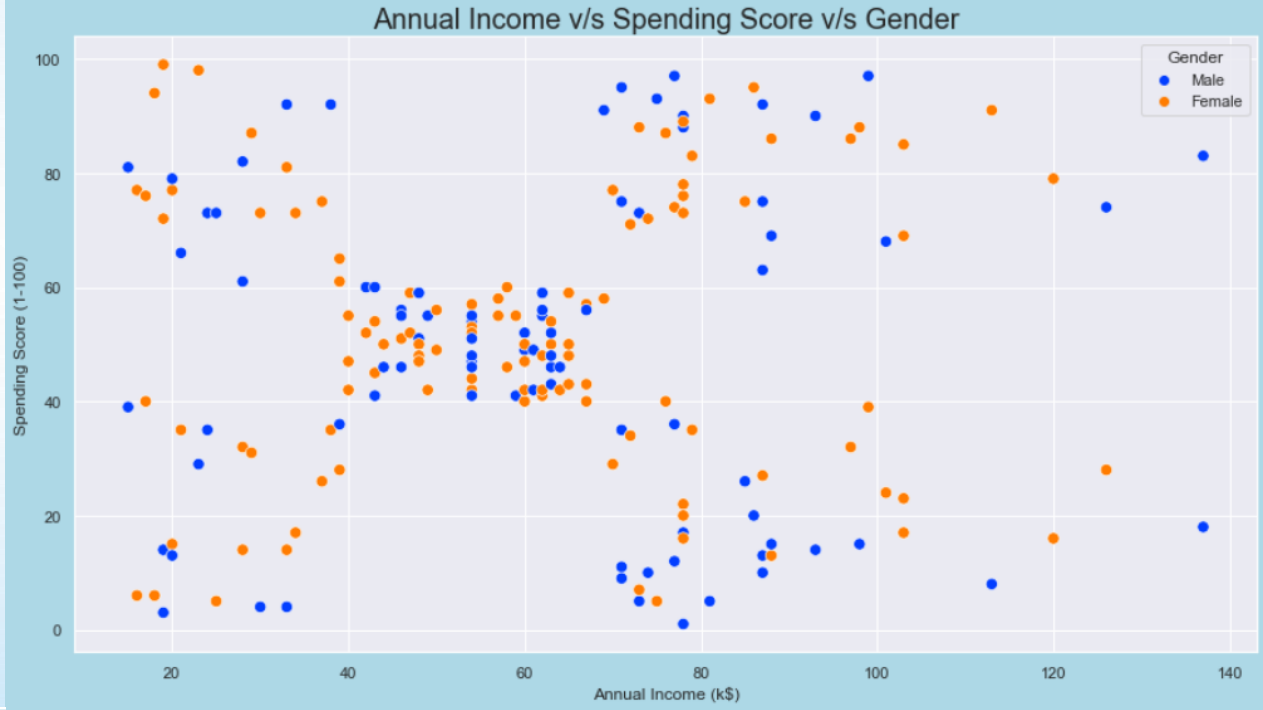


- BELOW PLOTS SHOWS THE DISTRIBUTION OF NUMERICAL FEATURES IN THE DATASET AND THERE IS NO OUTLIERS IN ANY FEATURES



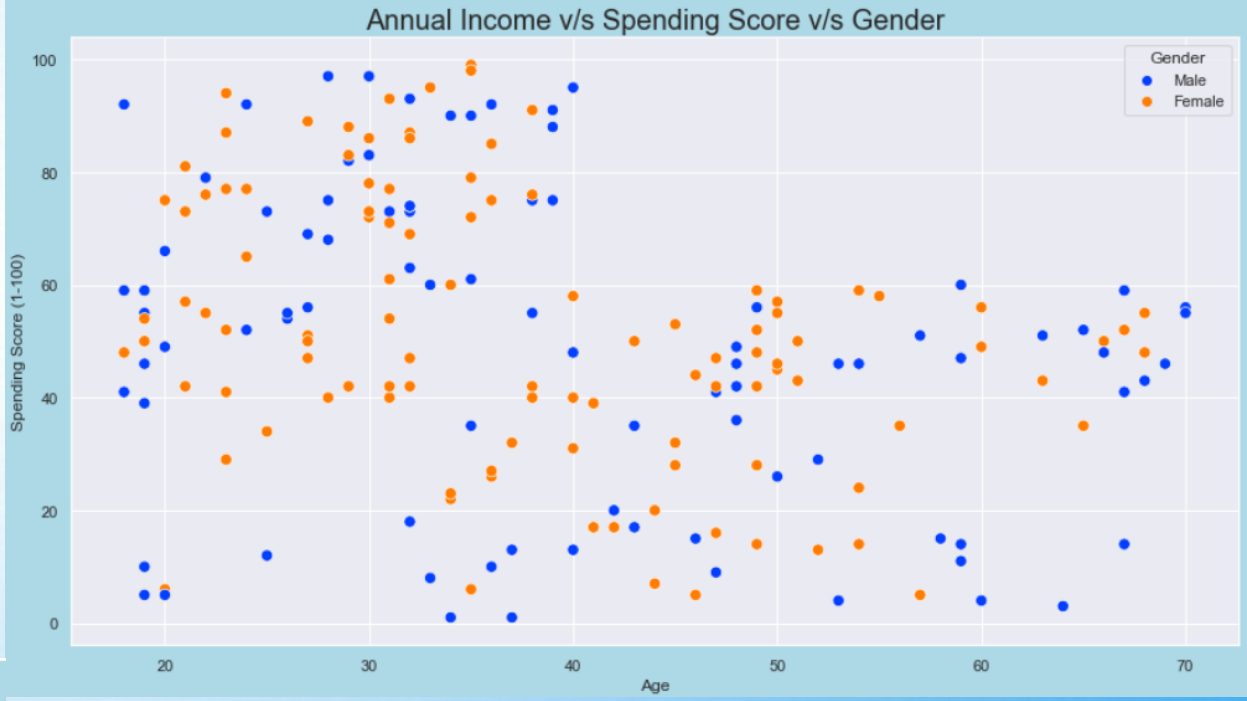
➢ BIVARIANT ANALYSIS

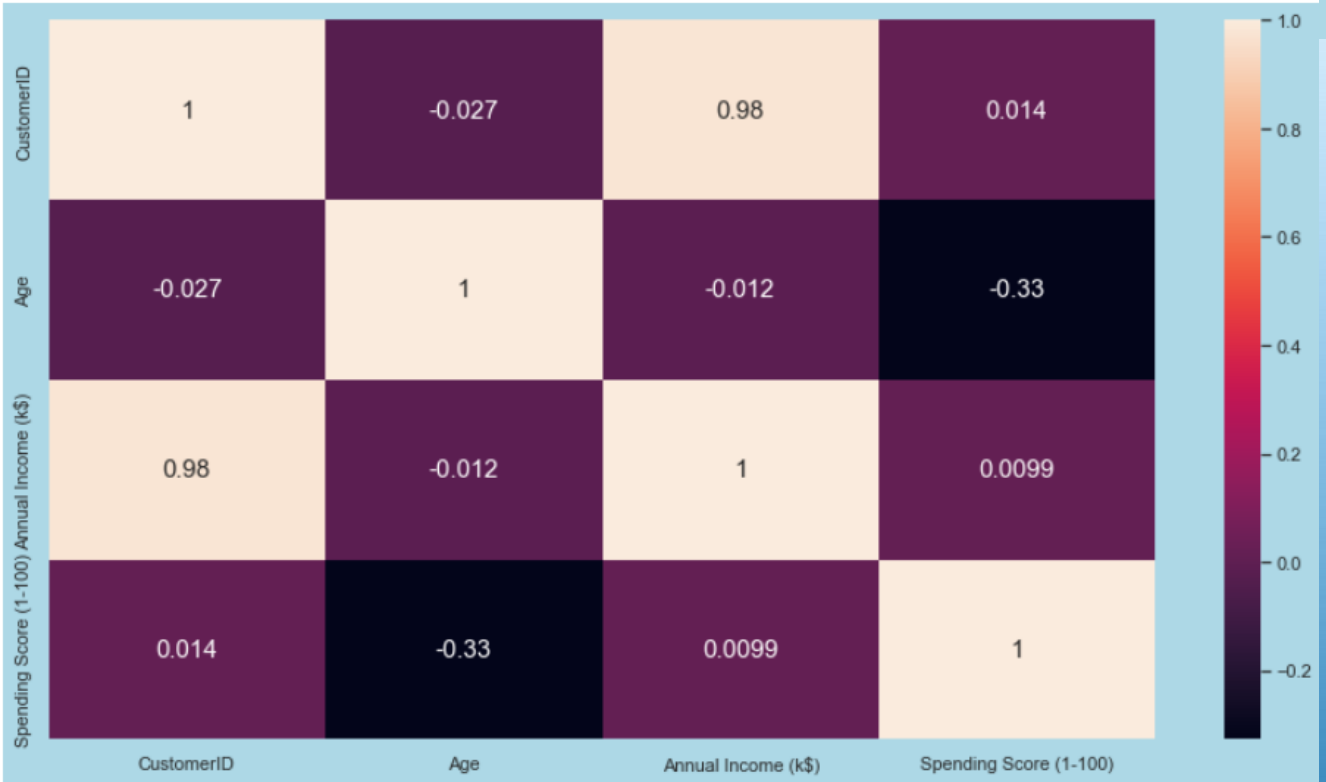
➢ HERE WE CAN SEE HOW THE ANNUAL INCOME AND SPENDING SCORE IS RELATED



➢ HERE WE CAN SEE HOW THE AGE AND ANNUAL INCOME RELATED

- HERE WE CAN SEE HOW THE AGE AND SPENDING SCORE IS RELATED





- WE CAN SEE HOW THE CORRELATION EXISTS BETWEEN ALL THE FEATURES

➢FEATURE ENGINEERING

➢FEATURE TRANFORMATION

- TO TRANSFORM ALL THE FEATURES INTO NUMERICAL DATATYPE
- LABEL ENCODING TECHNIQUE IS USED FOR FEATURE TRANSFORMATION

➢FEATURE SCALING

- TO GET ALL THE FEATURE INTO SIMILAR RANGE
- BUT IN THIS DATASET ALL THE FEATURES ARE IN SIMILAR RANGE SO NO NEED TO DO FEATURE SCALING FOR THIS DATASET

➢FEATURE SELECTION

- THERE ARE ONLY FEW FEATURES PRESENT IN THE DATASET AND ALSO ALL ARE RELEVANT EXCEPT ID FEATURE WE CAN DROP ID COLUMN

➢MODEL BUILDING

➢ EDA OBSERVATIONS

- THERE IS NO ANY DEPENDENT COLUMN PRESENT IN THE DATASET
- FROM THE DATA WE CAN SEE SOME GROUPS ARE AVAILABLE IN THE VISUALIZATION

➢ ALGORITHM SELECTION

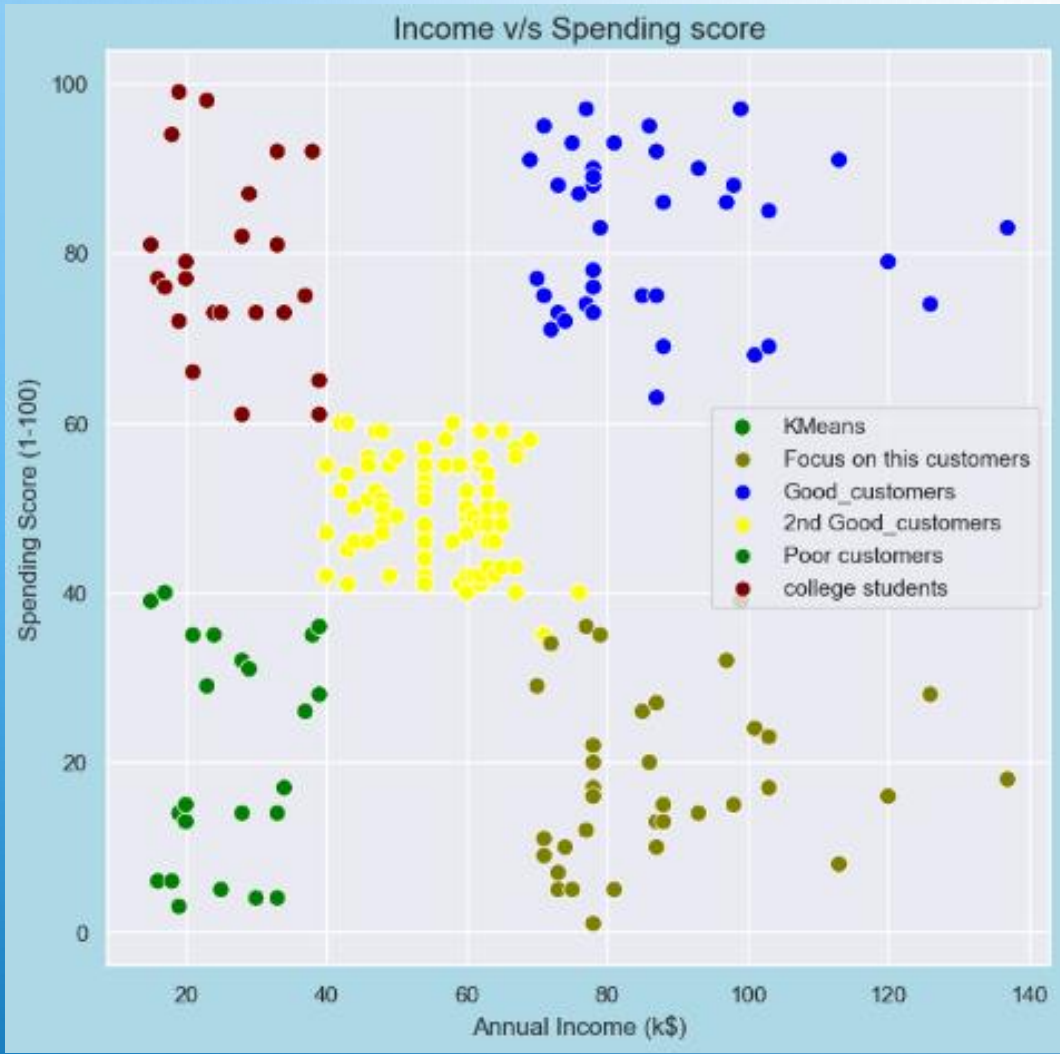
- THE ALGORITHM SHOULD BE UNSUPERVISED LEARNING ALGORITHM
- AS THE OUTLIERS ARE NOT PRESENT IN THE FEATURES SO ANY ALGORITHM CAN BE SELECTED
- WE CAN TAKE KMEANS BECAUSE HERE THERE IS NO OUTLIERS SO KMEANS IS GOOD TO CLUSTER FOR THIS TYPE OF DATA BUT TO COMPARE WHICH ONE IS GIVING BETTER PERFORMANCE KMEANS, AGGLOMERATIVE AND DBSCAN ALGORITHMS ARE SELECTED

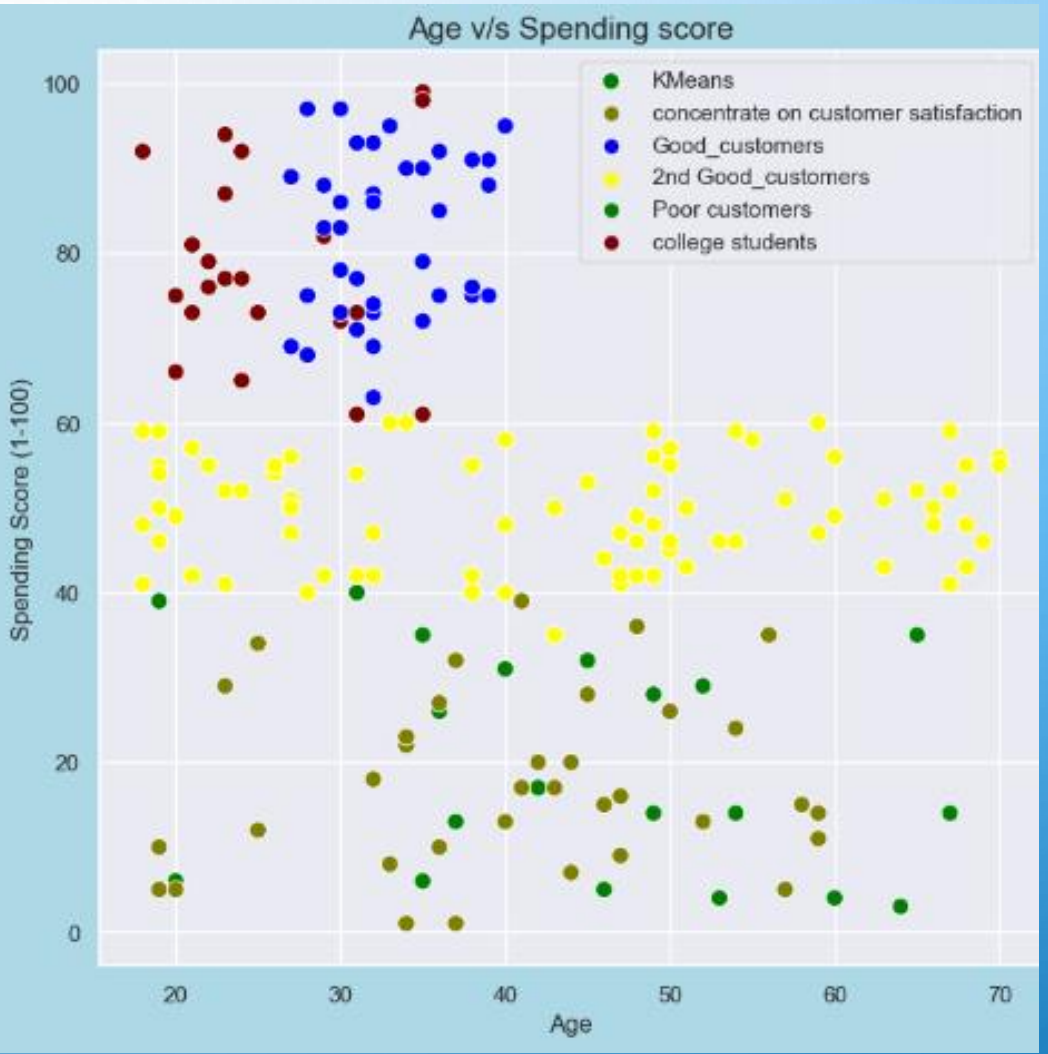
➢ MODEL PERFORMANCE AND EVALUATION RESULTS

	Algorithms	Silhouette score
0	Kmeans	0.444067
1	Agglomerative	0.439975
2	DBSCAN	0.255104

➢ FROM THE TABLE WE CAN CONCLUDE THAT KMEANS WITH FEATURE REDUCTION IS GIVING GOOD PERFORMANCE COMPARED TO OTHER ALGORITHMS

- FROM THE BELOW VISUALIZATION WE CAN SEE HOW THE GROUPS ARE MADE USING KMEANS ALGORITHM





➢ CONCLUSION

- **CLUSTER 1: HIGH INCOME / LESS SPENDING SCORE / HIGH AGE**
  - **NEED TO FOCUS ON CUSTOMER SATISFACTION TO TARGET ON THIS CUSTOMERS**
- **CLUSTER 2: HIGH INCOME / HIGH SPENDING SCORE / MEDIUM AGE**
  - **THESE ARE GOOD CUSTOMERS WHO SATISFIED WITH THE SERVICE**
- **CLUSTER 3: AVERAGE INCOME / AVERAGE SPENDING SCORE / AGE IS SPREADED THROUGHOUT**
  - **THESE ARE ALSO GOOD CUSTOMERS WHO SATISFIED WITH THE SERVICE SPENDING ACCORDING TO THERE INCOME**
- **CLUSTER 4: LOW INCOME / LOW SPENDING SCORE / MEDIUM AGE**
  - **THESE ARE SPENDING LESS BECAUSE OF LESS INCOME**
- **CLUSTER 5: LESS INCOME / HIGH SPENDING SCORE / YOUNG AGE**
  - **THESE ARE STUDENTS SO NEED TO ATTRACT BY FRIENDLY ENVIRONMENT**
- **THE KMEANS IS GIVING GOOD PERFORMANCE AND GOOD CLUSTERS**
- **THE CLUSTERS ARE GOOD TO MAKE DECISIONS**

THANK YOU