Final 'Presentation' { [Identifying Valuable Insights for Marketing Decision Based on Shopping Mall Dataset]

<Group 13: Yifan Zang, Harshdeep Singh, Shubh Goyal, Reny Marinez, Tapan Uchil >

}

Table Of 'Content' {

}

01 {

[Introduction to Dataset and Project]

}

Introduction {

General Research Interest: We aim to explore how age, gender and annual income affecting the spending score of customers sampled from a shopping mall population. Our analysis aims to extract valuable insights using segmentation for market basket analysis.

Dataset: Shopping Mall Customer Segmentation Dataset contains 15079 unique values. It includes basic customer data such as Customer ID, age, gender, annual income, and spending score-which is assigned based on customer behavior and purchasing data. The goal is to help a supermarket mall owner understand their customers better and identify target customers who are likely to converge, and therefore provide insights to the marketing team for strategic planning

} <DataSource:>https://www.kaggle.com/datasets/zubairmustafa/shopping-mall-customer-segmentation-data?resource=download

Dataset Overview using Dashboard

Definition of Spending Score

- Spending score is a composite variables that is based on factors including amount of money spend, purchase frequency,membership renewal rate and order return rates.

02 {

[General Research Interest]

We would like to explore how age, gender or annual income affects the spending score of customers and extract valuable insights for marketing decision

}

Data Analysis tasks

Technique and Tools: Kaggle API, Basic Statistic (Median, Mean, Percentile,Percentage)

Packages Utilized: Pandas and NumPy

```
# Call the function to initialize Kaggle API configuration
init_on_kaggle()
```

```
# Get in dataset using API. Data Source: https://www.kaggle.com/datasets/zubairmustafa/shopping-mall-customer-segmen
# The dataset Identifier is zubairmustafa/shopping-mall-customer-segmentation-data
!kaggle datasets download -d zubairmustafa/shopping-mall-customer-segmentation-data -p ~/Downloads --force
!unzip -o ~/Downloads/shopping-mall-customer-segmentation-data.zip -d ~/Downloads
import pandas as pd

data = pd.read_csv("~/Downloads/Shopping Mall Customer Segmentation Data .csv")
print(data.head())
```

```
Dataset URL: https://www.kaggle.com/datasets/zubairmustafa/shopping-mall-customer-segmentation-data
License(s): CC0-1.0
Downloading shopping-mall-customer-segmentation-data.zip to /Users/yfzwork/Downloads
  0%|                                               | 0.00/426k [00:00<?, ?B/s]
100%|██████████████████████████████████████| 426k/426k [00:00<00:00, 5.26MB/s]
Archive:  /Users/yfzwork/Downloads/shopping-mall-customer-segmentation-data.zip
  inflating: /Users/yfzwork/Downloads/Shopping Mall Customer Segmentation Data .csv
                            Customer ID  Age  Gender  Annual Income  \
0  d410ea53-6661-42a9-ad3a-f554b05fd2a7   30    Male         151479
```

SetUp API

CleanUp dataset by removing Null values and Customer ID

	Customer ID	Age	Gender	Annual Income	Spending Score
0	d410ea53-6661-42a9-ad3a-f554b05fd2a7	30	Male	151479	89
1	1770b26f-493f-46b6-837f-4237fb5a314e	58	Female	185088	95
2	e81aa8eb-1767-4b77-87ce-1620dc732c5e	62	Female	70912	76
3	9795712a-ad19-47bf-8886-4f997d6046e3	23	Male	55460	57
4	64139426-2226-4cd6-bf09-91bce4b4db5e	24	Male	153752	76

Cleaned Up!

Total records for 'Customer ID': 15079

	Age	Gender	Annual Income	Spending Score
0	30	Male	151479	89
1	58	Female	185088	95
2	62	Female	70912	76
3	23	Male	55460	57
4	24	Male	153752	76

Basic Statistics for EDA and Brainstorm of Subquestions

```
Analysis of 'Spending Score' for Age Group: Under 18
  Median: 50.5
  Average (Mean): 50.097938144329895
  Range: 99 (Min: 1, Max: 100)
  Percentiles:
0.25    24.25
0.50    50.50
0.75    74.00
0.90    92.70
0.95    96.00
0.99    98.07
Name: Spending Score, dtype: float64
```

```
Analysis of 'Spending Score' for Age Group: 30-45
  Median: 50.0
  Average (Mean): 50.39475388601036
  Range: 99 (Min: 1, Max: 100)
  Percentiles:
0.25     26.0
0.50     50.0
0.75     76.0
0.90     90.0
0.95     95.0
0.99    100.0
Name: Spending Score, dtype: float64
```

```
Analysis of 'Spending Score' for Age Group: 18-30
  Median: 51.0
  Average (Mean): 51.04276985743381
  Range: 99 (Min: 1, Max: 100)
  Percentiles:
0.25     26.0
0.50     51.0
0.75     76.0
0.90     91.0
0.95     96.0
0.99    100.0
Name: Spending Score, dtype: float64
```

```
Analysis of 'Spending Score' for Age Group: 45-60
  Median: 51.0
  Average (Mean): 51.19832312157369
  Range: 99 (Min: 1, Max: 100)
  Percentiles:
0.25     27.0
0.50     51.0
0.75     76.0
0.90     91.0
0.95     96.0
0.99    100.0
Name: Spending Score, dtype: float64
```

Screenshot example:

We do not see significant differences of spending score among different age group based on median,mean,percentile,etc

Similar situations apply for different gender and income groups.

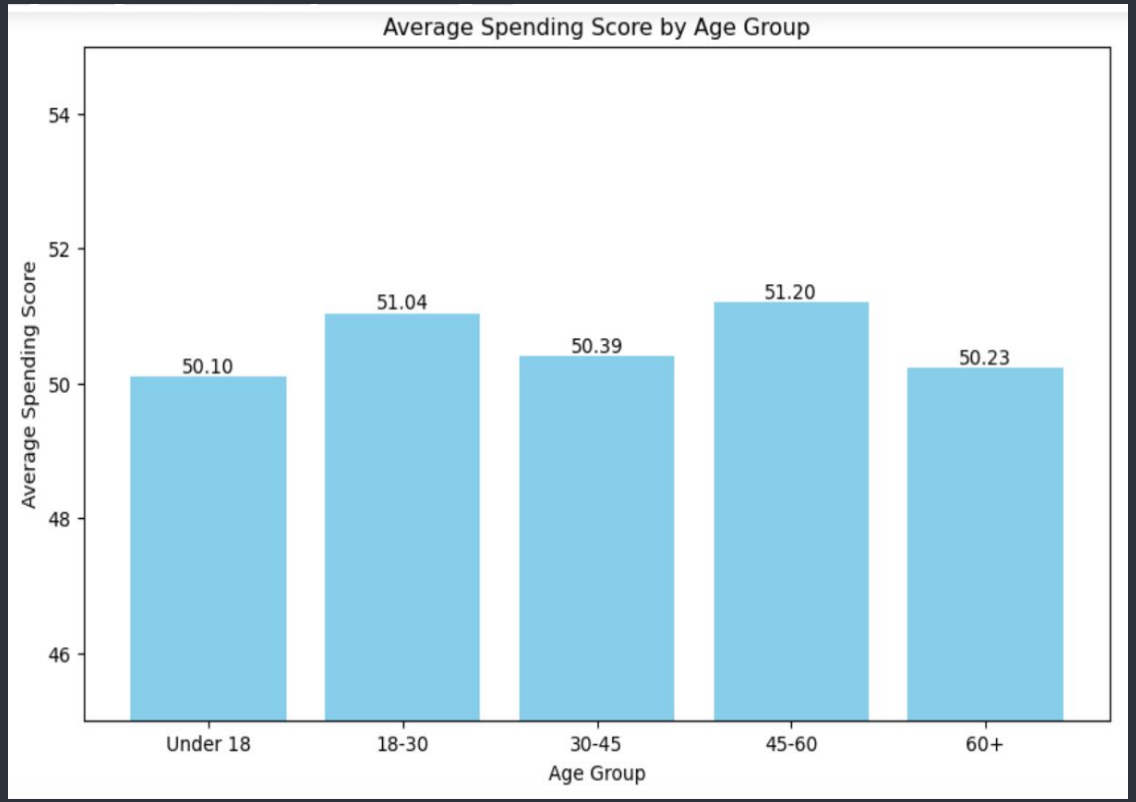
Then we would to use visualization technique for extracting valuable insights that might be obvious.

03-A {

[Sub-question:
What segmentations are valuable?]

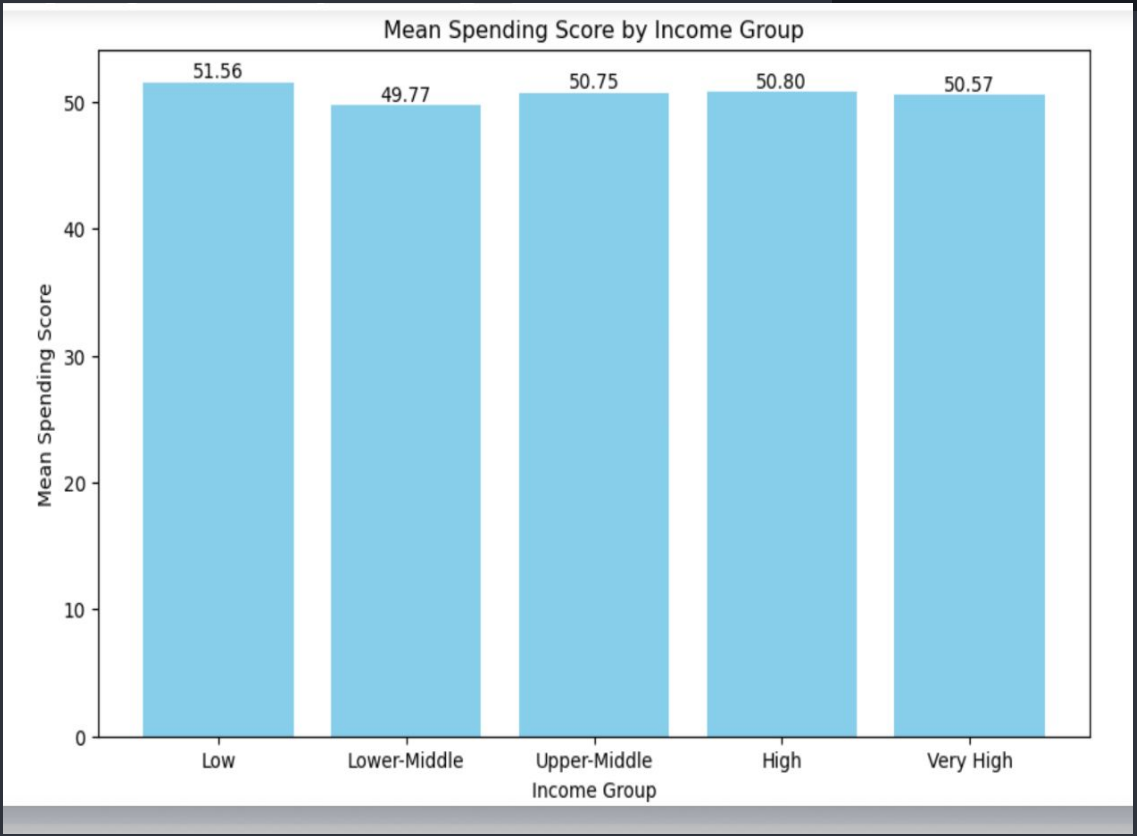
}

Average Spending Score by Age Group



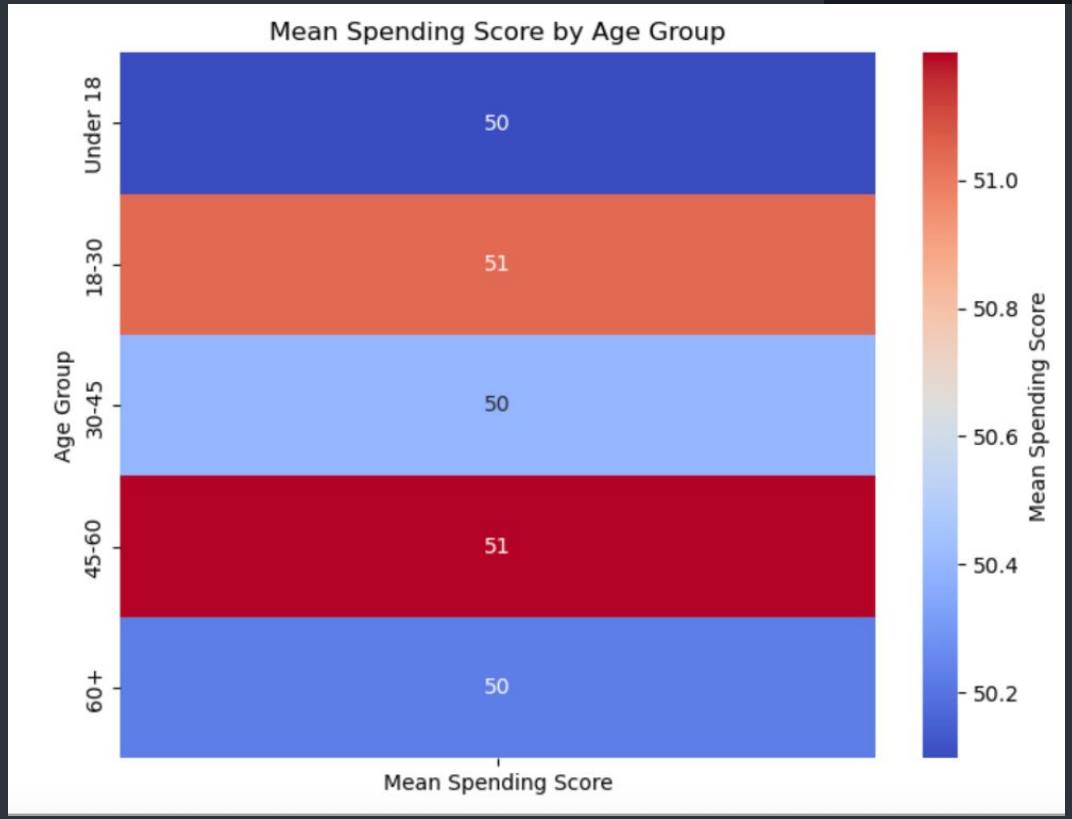
- The average spending scores across different age groups are relatively similar
- This suggests that while age may have some impact on spending behavior, in this sample, the difference is not very pronounced.
- The age groups of 18-30 and 45-60 show slightly higher average spending scores than the other groups
- Despite varying ages, the spending scores remain mostly around the 50 mark

Mean Spending Score by Income Group



- **Contrary to what might be expected, the Low income group has the highest mean spending score**
- **The spending scores for the lower-middle, upper-middle, high, and very high income groups hover around 50, showing less variation compared to the low income group**
- **More uniform spending pattern amongst these groups**

Mean Spending Score by Age Group

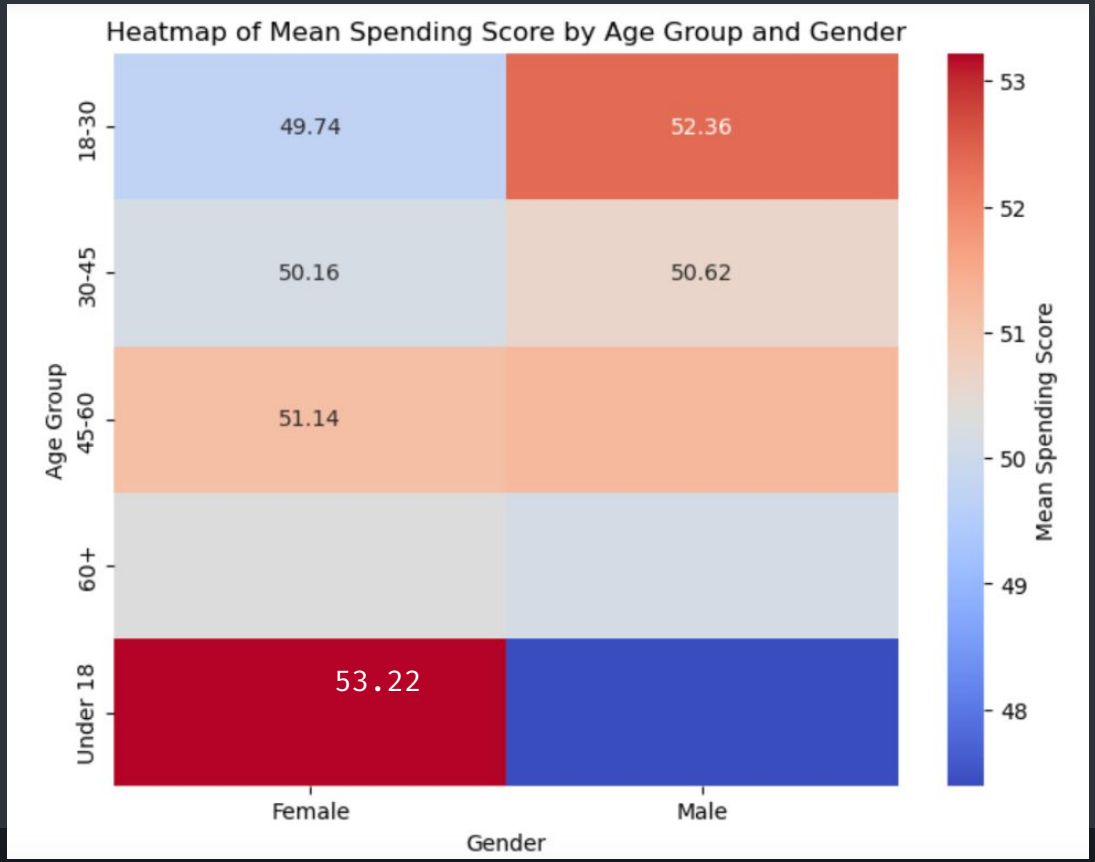


- This graph shows that the age groups 18-30 and 45-60 have slightly higher spending scores compared to the other groups, indicated by a the orange and red shade
- The age groups under 18, 30-45, and 60+ have a spending score of 50, represented by a light blue shade

Average Spending Score By Age Group And Gender

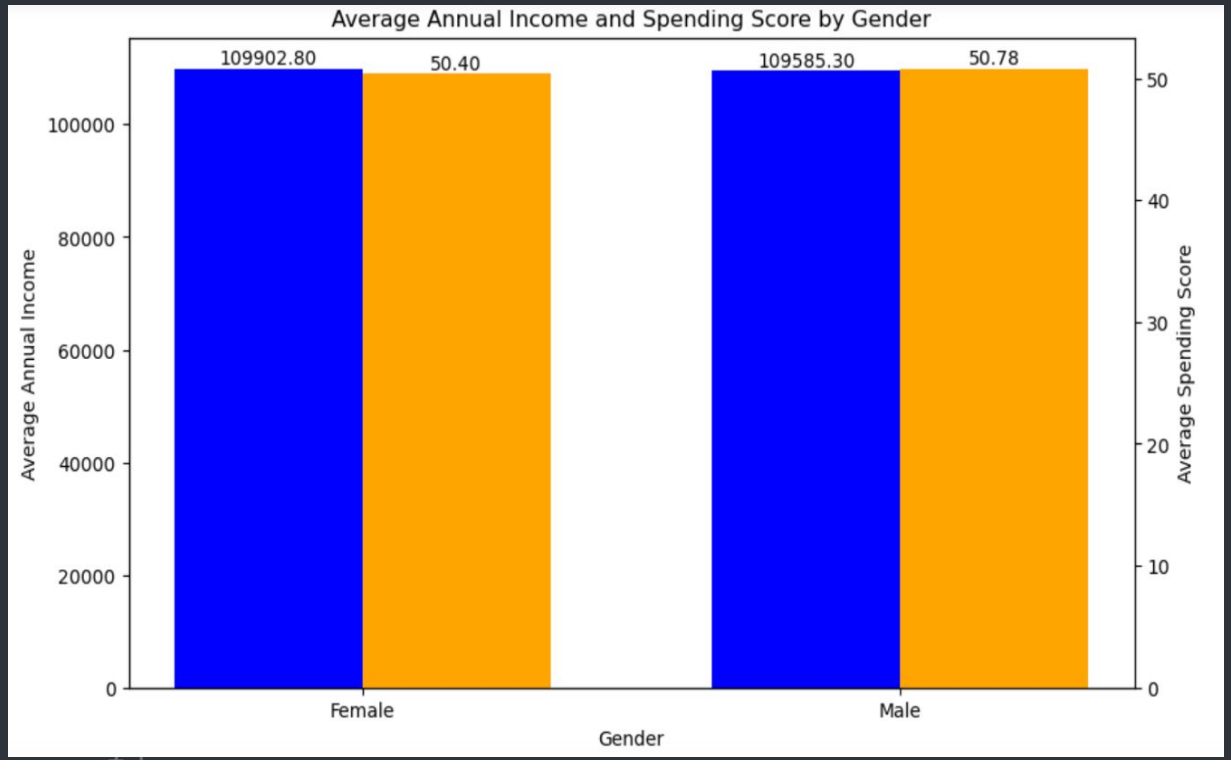
- **Young Female customers Spend More. Females under 18 have a higher average spending score than any other group.**
- **Males in the 30-45 age group show a higher spending score compared to females in the same group**
- **For those 60 and older, the spending scores between genders are very close**

Another Visual Technique:
Heatmap of Mean Spending Score by Age Group and Gender



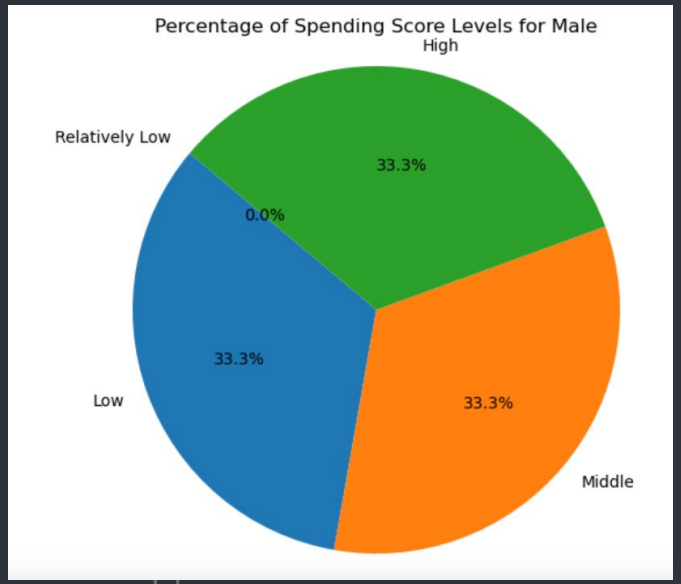
- There is a notable gender difference in spending behavior, particularly in the under 18 female group and 18-30 male group as they shows significantly higher average spending score
- For the age groups 30-45, 45-60, and 60+, the spending scores are relatively similar across genders.
- Both genders in the 45-60 age group show elevated spending scores compared to other age groups

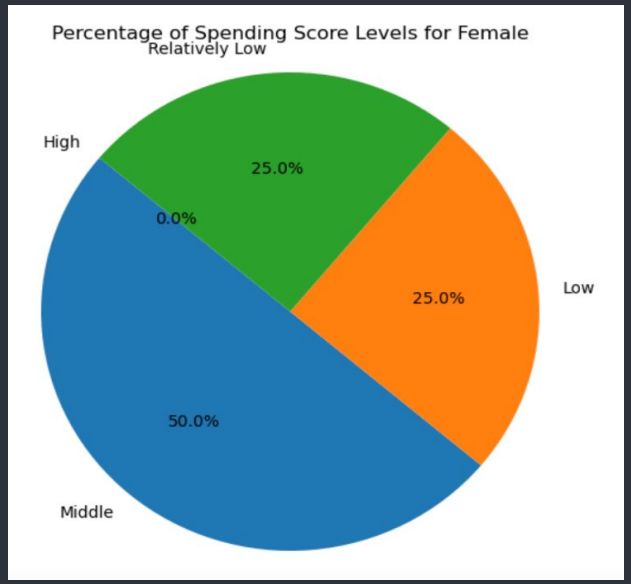
Spending Score by Gender in Relation to Annual Income



- **Gender and annual income may not significantly differentiate spending habits in this sample.**
- **Such phenomenon occurs due to evolving societal norms, individual differences,the product offerings at the mall, potential sample bias, and nuanced data interpretation.**

Percentage of Spending Score Level by Gender





Score Category

Low:	0-25
Relatively Low:	25-50
Middle:	50-75
High:	75-100

For Male category, the distribution for spending score is relatively balanced with each account for 33% for relatively low group (25-50),middle(50-75) and high (75-100).

For female category, we see 50% of female consumers demonstrate middle spending score.

03-A

[Can we find Statistically Significant relationship between
(1)Age vs spending score
(2)Annual Income vs Spending Score ?]

‘01’ Technique: Simple linear Regression{

Question

Is there a significant statistical relationship between a customer's age and their spending score in the mall?

Hypothesis

Ho: There is no significant statistical relationship between age and spending score.

Ha: There is a statistically significant relationship between age and spending score.

}

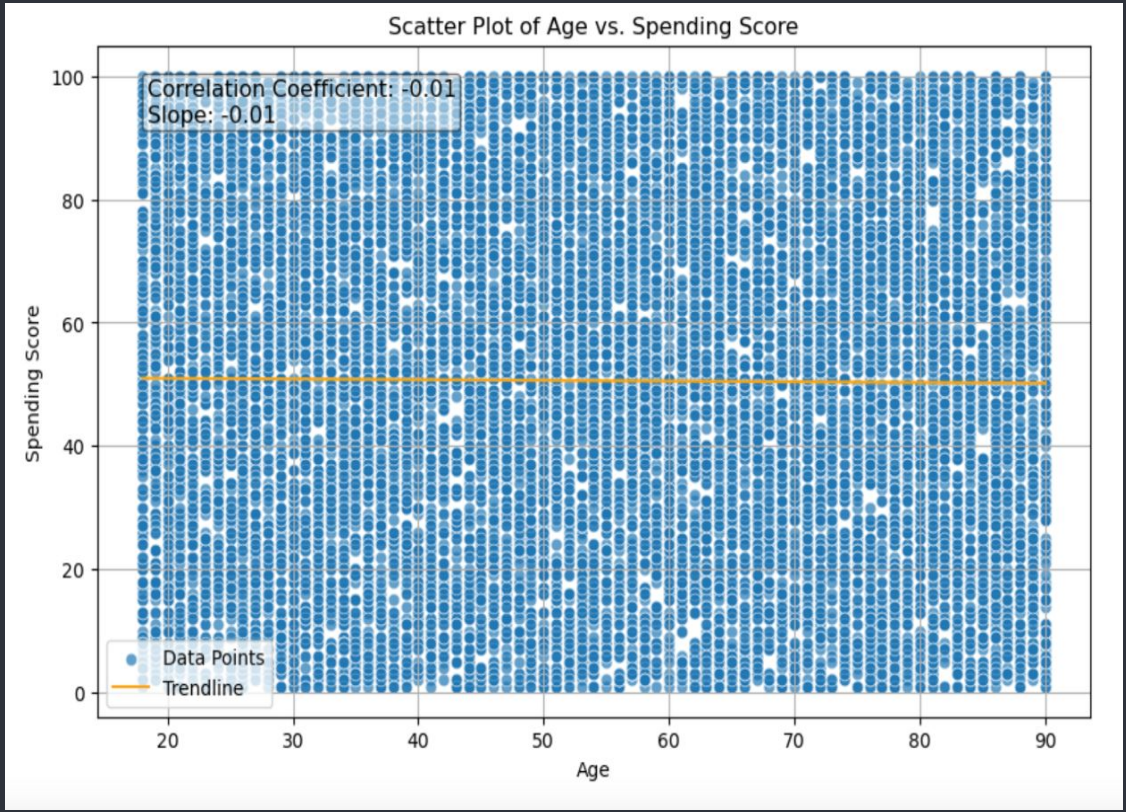
Simple Linear Regression Output and Interpretation Age vs Spending Score

```
                            OLS Regression Results
==============================================================================
Dep. Variable:          Spending Score   R-squared:                       0.000
Model:                             OLS   Adj. R-squared:                  0.000
Method:                  Least Squares   F-statistic:                     1.027
Date:                 Sat, 11 May 2024   Prob (F-statistic):              0.311
Time:                         07:03:41   Log-Likelihood:                -72028.
No. Observations:                15079   AIC:                         1.441e+05
Df Residuals:                    15077   BIC:                         1.441e+05
Df Model:                            1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const         51.1998      0.644     79.469      0.000      49.937      52.463
Age           -0.0112      0.011     -1.013      0.311      -0.033       0.010
==============================================================================
Omnibus:                    11371.677   Durbin-Watson:                   2.004
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              879.408
Skew:                          -0.002   Prob(JB):                    1.09e-191
Kurtosis:                       1.817   Cond. No.                        160.
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
Correlation Coefficient between Age and Spending Score: -0.008251286264616638
```

1. **R-squared (0.000)**: An R-squared of 0.000 suggests that the model explains none of the variability in the spending scores around its mean. Essentially, "Age" does not explain any variation in "Spending Score."
2. **P-value for Age (0.311)**: This high p-value suggests that Age is not statistically significant in predicting Spending Score

Age vs Spending Score
Scatter Plot with Trend Line



1. Trend line is near horizontal with slope at -0.01
2. Points of Spending Score are scattered around with obvious pattern

No Trend !

Sub-Question '01' {

Question

Is there a significant statistical relationship between a customer's annual income and their spending score in the mall?

Hypothesis

Ho: There is no significant statistical relationship between age and spending score.

Ha: There is a statistically significant relationship between age and spending score.

}

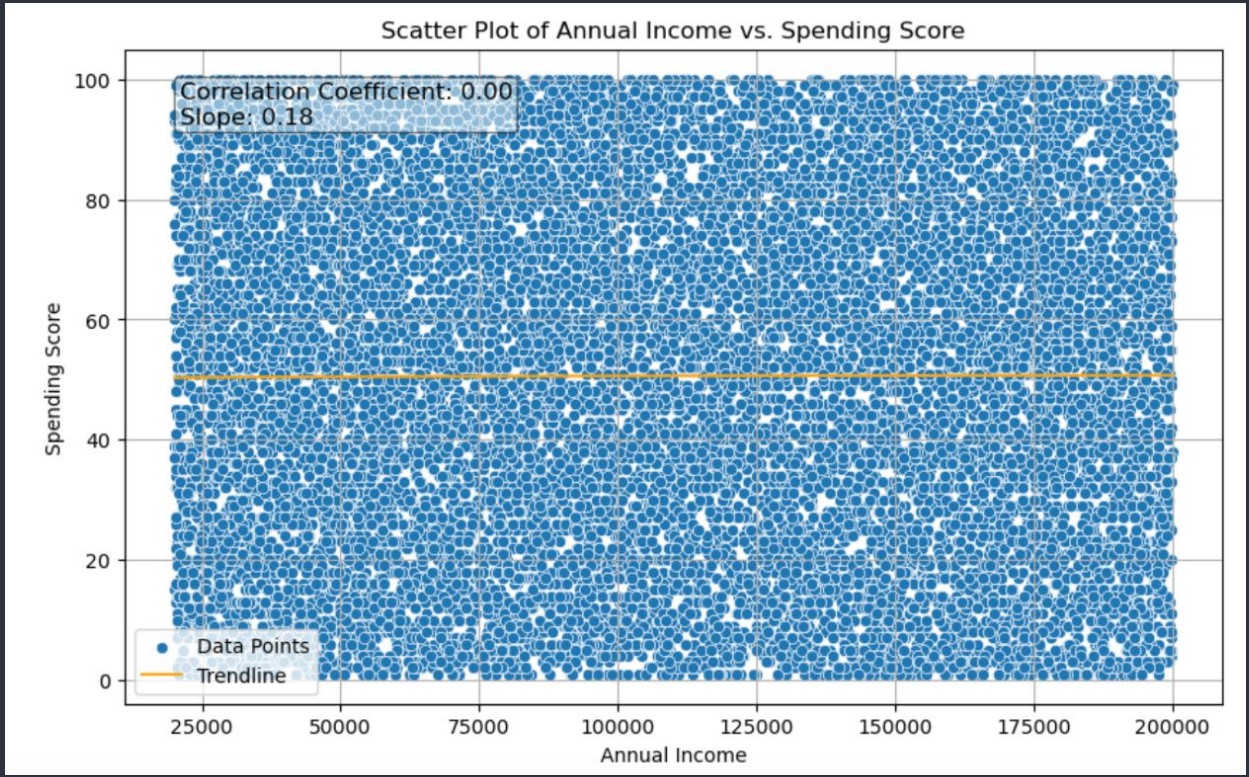
Simple Linear Regression Output and Interpretation
Annual Income vs Spending Score

```
                            OLS Regression Results
==============================================================================
Dep. Variable:          Spending Score   R-squared:                       0.000
Model:                            OLS   Adj. R-squared:                 -0.000
Method:                 Least Squares   F-statistic:                     0.2039
Date:                Sat, 11 May 2024   Prob (F-statistic):              0.652
Time:                        07:03:43   Log-Likelihood:                -72028.
No. Observations:               15079   AIC:                         1.441e+05
Df Residuals:                   15077   BIC:                         1.441e+05
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const         48.5456      4.537     10.700      0.000      39.653      57.438
Annual Income  0.1786      0.395      0.452      0.652      -0.597       0.954
==============================================================================
Omnibus:                    11344.850   Durbin-Watson:                   2.004
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              878.992
Skew:                          -0.002   Prob(JB):                    1.35e-191
Kurtosis:                       1.817   Cond. No.                         224.
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
Correlation Coefficient: 0.003677753432072137
```

1. R-squared (0.000): The model does not effectively explain the changes in Spending Score based on Annual Income
2. Coefficients: (0.1786): This coefficient indicates a minimal impact for annual income on spending score
3. P-value for Annual Income (0.652): This high p-value indicates that the Annual Income coefficient is not statistically significant, supporting the evidence that Annual Income does not effectively predict Spending Score.

Annual Income vs Spending Score Scatter Plot with Trend Line



1. Trend line is near horizontal with slope at 0.18
2. Points of Spending Score are scattered around with no obvious pattern
3. Correlation coefficient is practically 0.

No Trend !

05 {

[Methodology Used]

}

Methodology and Technique Recap {

Data Exploration: Identifying variables relevant to the data and calculating basic statis
including Mean, Median, Percentile to understand the dataset.
Tools and packages used are: Kaggle API,Pandas,Numpy

Data Cleaning: Remove missing values and customer ID;
Log Scale of Annual Income for Regression analysis

Statistical Analysis: Hypothesis Testing, Simple Linear Regression and Graph Evaluation
Packages involved: Statsmodels.api

Data Visualisation: Bar Plot, Pie Chart, HeatMap, Scatterplot with TrendLine
Packages we use: Seaborn,Matplotlib

Marketing Analysis Methodology: Segmentation, Market Basket Analysis

06 {

[Key Findings]

}

Valuable Segmentation for Marketing{

Findings of Age Group

- The age groups of 18-30 and 45-60 show slightly higher average spending scores than the other groups

Recommendation

< We recommend develop targeted marketing strategy for these 2 age groups due to their high spending scores. Also consider exploring other factors such as shopping preferences for such age groups>

}

Gender-Based Findings{

Findings

< Males tend to have a slightly higher spending score than females. The distribution of spending scores across gender also shows that both males and females have a balanced spending behavior

>

Recommendations

< Instead of relying solely on gender-based segmentation, we propose personalized marketing campaigns by considering more factors such as macroeconomic condition, social environment and trendy products on market>

}

Income-Related Findings {

Findings

- Contrary to what might be expected, the Low income group has the highest mean spending score
- The spending scores for the lower-middle, upper-middle, high, and very high income groups hover around 50, showing less variation compared to the low income group

Recommendations

< We suggest identifying product needs and service preference for low-income groups. For other income groups, the marketing activities should consider a multifaceted forces of demographic characters, psychographic traits, and behavioral pattern to ensure the effectiveness of marketing campaign.>

}

Conclusion{

We aim to explore how age, gender and annual income affecting the spending score of customers sampled from a shopping mall population.

However, the variables in dataset might not be effective on predicting a certain trend on a general sense.

It is the detailed segmentation based on one or two factors that drive valuable marketing campaign.

Thank You!