|  |  |  |  |
| --- | --- | --- | --- |
| Team 5   |  | | --- | | PROJECT REPORT: detail of persons hands with scissors, markers, workingPredictive Analysis of Customer Spending Power (Customer Segmentation) | |  |  |
|  |  |  |

TABLE OF CONTENTS

[Predictive Analysis of Customer Spending Power (Customer Segmentation) 3](#_Toc15397941)

[Strategic Highlights / Business Case 3](#_Toc15397942)

[Objective 3](#_Toc15397943)

[Data Collection 3](#_Toc15397944)

[Data Preparation 4](#_Toc15397945)

[Data Exploration & Visualization 4](#_Toc15397946)

[Age Grouping 9](#_Toc15397947)

[Model Building 10](#_Toc15397948)

[I. OLS with Forward Selection 10](#_Toc15397949)

[Results 11](#_Toc15397950)

[II. Bayesian Linear Regression with PyMC3 12](#_Toc15397951)

[Conclusions 15](#_Toc15397952)

[FINANCIAL STATEMENTS 16](#_Toc15397953)

[Statement of Financial Position 16](#_Toc15397954)

[Statement of Comprehensive Income (Profits and Losses) 16](#_Toc15397955)

[Statement of Changes in Equity 16](#_Toc15397956)

[Statement of Cash Flows 16](#_Toc15397957)

[References 17](#_Toc15397958)

# Predictive Analysis of Customer Spending Power (Customer Segmentation)

## Strategic Highlights / Business Case

*Through extensive examination of the trends in spending score, predict the spending power of the target customers.*

**Customer segmentation** is the practice of dividing a customer base into groups of individuals that are similar in specific ways relevant to marketing, such as age, gender, interests and spending habits.

**Customer Segmentation for Target Marketing:** Companies employing customer segmentation operate under the fact that every customer is different and that their marketing efforts would be better served if they target specific, smaller groups with messages that those consumers would find relevant and lead them to buy something. Companies also hope to gain a deeper understanding of their customers' preferences and needs with the idea of discovering what each segment finds most valuable to more accurately tailor marketing materials toward that segment.

## Objective

1. Analyze the dataset to understand the dependency of spending score on available factors: Age, Annual Income, and Gender.
2. Create 2 models to predict spending score of the target customers (in the available dataset).
3. Comparative analysis of the two models (Ordinary Least Squares (OLS) with Forward Selection, and Bayesian Linear Regression with PyMC3)
4. Draw conclusion for target customers from the available customer dataset.

## Data Collection

Through membership cards, some basic data about customers like Customer ID, age, gender, annual income and spending score are gathered for a shopping mall’s customers.

(The data in our case study is simulated)

***Spending score is a score assigned to the customer based on certain defined parameters like customer behavior and purchasing data.***

|  |
| --- |
| ***Problem Statement: Understand the likelihood of buying or predicting the customers who are more likely to converge [Target Customers] so that marketing strategies could be designed to cater to those target customers.*** |

## Data Preparation

**Dataset Details:**

* **CustomerId:** Customer's unique ID,
* **Gender:** Customer's gender
* **Age:** Customer's age
* **Annual Income:** Customer's annual income
* **Spending Score:** A score, out of 100, to rate customer's behavior and money spent by the customer.

|  |  |
| --- | --- |
| **Independent Variables** | **Dependent Variable** |
| Age, Annual Income, Gender | Spending Score |

A screenshot of a cell phone

Description automatically generated

Figure 1: Snippet of Dataset

## Data Exploration & Visualization

Basic statistical details of the data set:

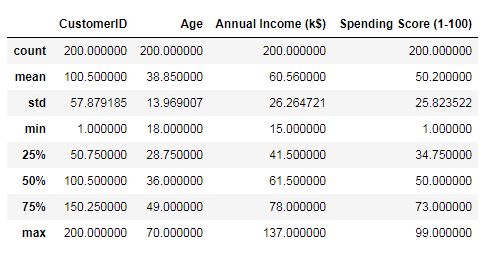


Figure 2: Basic Statistical Details of Dataset

Using the build in function, df.describe(), as shown above, we see that there are two hundred customers. For the category Age, we see that it ranges from 18 to 70 with a mean of 38.85 and a standard deviation of 13.969 years old. For the category Annual Income, we see that it ranges from $15k to $137k. It has an average of $60.56k with a standard deviation of $26.26k. Lastly, we see that the spending score ranges from 1 to 99 with a mean of 50.2 and standard deviation of 25.823.

**Histogram Plots:**

We plotted the histogram using matplotlib.pyplot. From the histogram plots we see that the dependent variables and independent variables are normally distributed. Therefore, we can use OLS regression analysis on this data.

A close up of a map

Description automatically generated

Figure 3: Histogram Plots

**Boxplot & Swarmplots**

A close up of a map

Description automatically generated

Figure 4: Boxplots and Swarmplots

**Summary observation on original dataset:**

We see that Age and Annual Income are not colinear. The correlation between the spending score and age and, spending score and annual income is low. Thus, it would be appropriate to conduct regression analysis on our dataset.

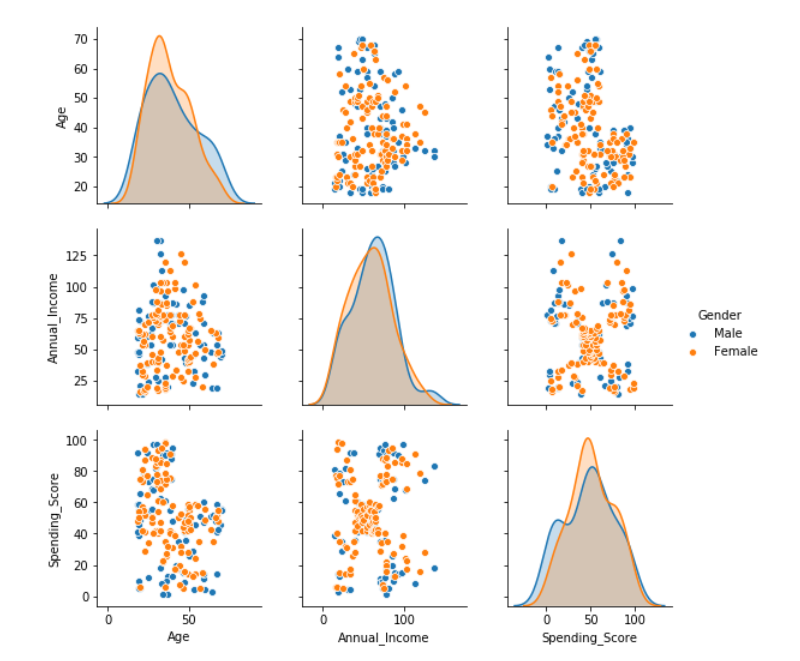
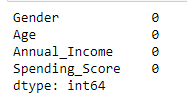


Figure 5: Pairplots on the Original Dataset

## Data Manipulation

Using the build in function, isnull(), we checked where there are missing data or bad data. We see that there are no null data in the data set

In the next step, we dropped CustomerID. It is just a unique id for each customer and therefore will not impact customer’s spending score.

Since Gender is a categorical string data, it needs to transformed before it can be used in data analysis. We mapped 0 for female customers and 1 for male customers

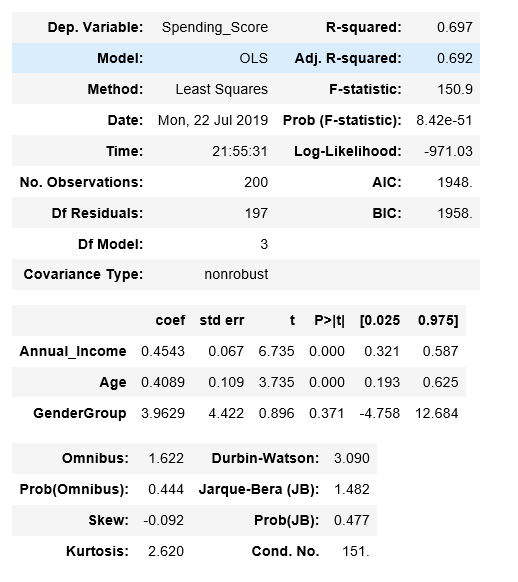


Figure 6: OLS Regression Results on Original Dataset

We see a low R-squared value of 0.697. p-value for GenderGroup is high (>0.05) showing that it is not a significant factor for the change in spending score. Therefore, we can drop GenderGroup from our model.

Other observations and summary include:

1. No clear linear relationship between the spending \_score and independent variables like annual\_income and age.

2. A single linear multivariate regression that models the relationship between the dependent variable and the independent variables returns a low R-squared at around 69.7%

3. An intercept should be excluded as it is assumed that having no annual income would suggest having no spending power/score.

## Age Grouping

We will divide the data into the age groups and try to run regression analysis for each age group. This is done in order to find out which age group has more significant effect on the spending score.

We split age into **6 Age Groups**: Below 25, 26 – 35, 36 – 45, 46 – 55, 56 – 65, Above 65.

A screenshot of a cell phone

Description automatically generated

Figure 7: Dataset snippet after age grouping

# Model Building

## OLS with Forward Selection

**Forward Selection**: Forward selection is a type of stepwise regression which begins with an empty model (dependent variable) and adds in the independent variables one by one. In each forward step, we add only the independent variable that maximizes R2 of the model.

We begin with only an intercept. We test the various variables that may be relevant, and the ‘best’ variable—where “best” is determined by a pre-determined criteria—is added to the model.

As the model continues to improve (per that same criteria) we continue the process, adding in one variable at a time and testing at each step. Once the model no longer improves with adding more variables, we stop the process.

As descried above we applied a forward elimination on the linear multivariate regressions on each age group to select the combinations of independent variables that yield highest R-squared.

Details on the final model:

# Model I - Results

**Results of regression analysis (OLS with Forward Selection):**

|  |  |  |
| --- | --- | --- |
| **Age Group** | **Regression Equation** | **R-squared** |
| Below 25 | Spending\_Score = 4.0296 \* Age - 0.6769 \* Annual\_Income  (T-value:0) (T-value:0) | 89% |
| 26 - 35 | Spending\_Score = 1.6762 \* Age + 0.1718 \* Annual\_Income  (T-value:0) (T-value:0.11) | 86% |
| 36 - 45 | Spending\_Score = 0.5828 \* Age + 0.2917 \* Annual\_Income  (T-value:0.13) (T-value:0.15) | 70% |
| 46 - 55 | Spending\_Score = 0.7401 \* Age  (T-value:0) | 83% |
| 56 - 65 | Spending\_Score = 0.5358 \* Age  (T-value:0) | 71% |
| Above 66 | Spending\_Score = 0.2173 \* Age + 0.6344 \* Annual\_Income  (T-value:0.01) (T-value:0.19) | 97% |

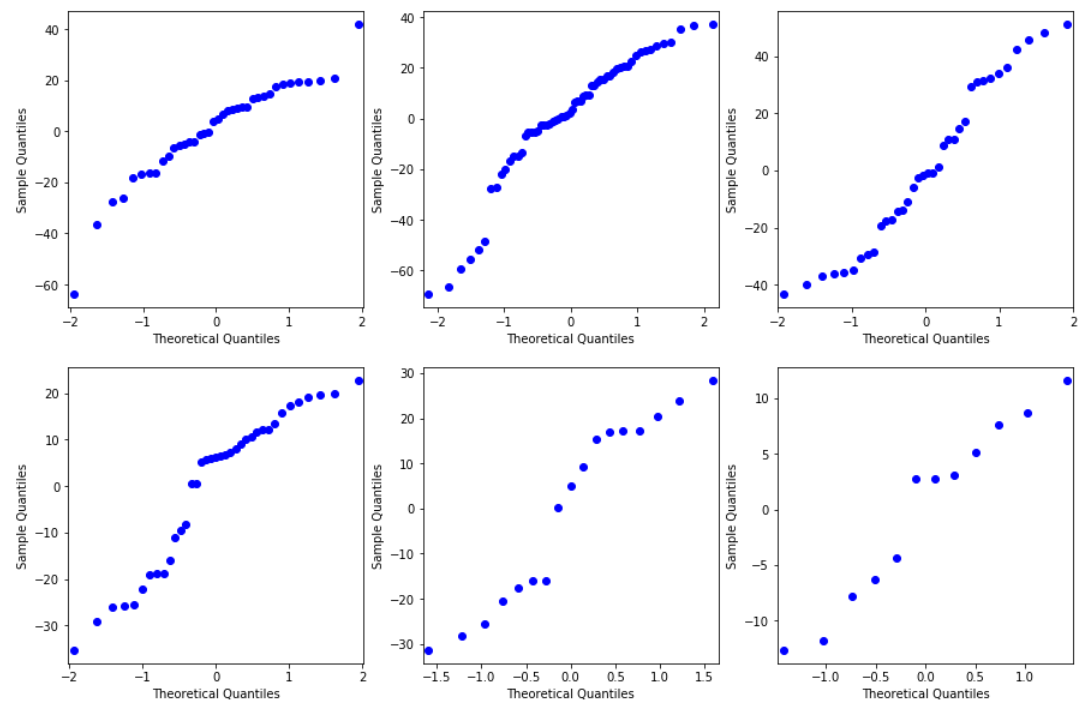


Figure 8: qqplot of Residuals for the 6 Age Groups (OLS)

From the above we can see that the residuals are along the diagonal, which shows normal distribution of residuals. The residuals are nearly normally distributed & centered on 0.

**Observations:**

1. R-squared for regression on each age group improve significantly.

2. the residuals are normally distributed.

3. Some independent variables have T-value higher than 5%, suggesting that they are not a significant determinant for the spending\_score. However, we argue that including those variables give us the best predicting power as combing them with other variables yield highest R-squared.

4. Annual\_Income is negatively correlated with the spending\_score for customers below 25.

5. Annual\_Income does not seem to have any relationship with the spending\_score for customers between 46 years old to 65 year old.

6. there are not many data for the age group above 56 (56-65 and above 66). Therefore, the regression results in these two groups could change significantly with more data collection.

## II. Bayesian Linear Regression with PyMC3

Now, we are going to introduce regression modelling in the Bayesian framework and carry out inference using the PyMC3 MCMC library.

In a Bayesian framework, linear regression is stated in a probabilistic manner. That is, we reformulate the above linear regression model to use probability distributions. In the Bayesian formulation we receive an entire probability distribution that characterises our uncertainty on the different β coefficients. The immediate benefit of this is that after taking into account any data we can quantify our uncertainty in the β parameters via the variance of this posterior distribution. A larger variance indicates more uncertainty.

**Plotting the heatmap to find out correlation between different parameters:**

A screenshot of a cell phone

Description automatically generated

Figure 9: Bayesian Regression - Heatmap

# Conclusions

Since our dataset is simulated and is not …

However, doing a 2 model comparison,

* For example, this is the List Bullet style.
* Here is another sentence formatted in List Bullet style.

You can find easy-to-use tools on the Insert tab, such as to add a hyperlink, insert a comment, or add automatic page numbering.

|  |
| --- |
|  |

View and edit this document in Word on your computer, tablet, or phone. You can edit text; easily insert content such as pictures, shapes, and tables; and seamlessly save the document to the cloud from Word on your Windows, Mac, Android, or iOS device.

# FINANCIAL STATEMENTS

## Statement of Financial Position

* Liabilities
* Statement of Financial Position
* Ownership Equity

## Statement of Comprehensive Income (Profits and Losses)

* Income
* Expenses
* Profits

## Statement of Changes in Equity

Well, it wouldn’t be an annual report without a lot of numbers, right? This section is the place for all those financial tables.

To get started with a table that looks just like the sample here, on the Insert tab, tap Table.

|  |  |  |  |
| --- | --- | --- | --- |
| **DESCRIPTION** | **REVENUE** | **EXPENSES** | **EARNINGS** |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |

## Statement of Cash Flows

* Operating
* Investing
* Financing

# References

* <https://searchcustomerexperience.techtarget.com/definition/customer-segmentation>