|  |
| --- |
| detail of persons hands with scissors, markers, workingPROJECT REPORT: Predictive Analysis of Customer Spending Power (Customer Segmentation) |

|  |  |  |
| --- | --- | --- |
| Team 5 |  |  |
| Anshita Varshney, Bhavnil Patel, Chongli Zhao, Dhairya Sheth, Jyoti Lnu, Ruofan Wu  Jyoti Lnu  Ruofan Wu |  |  |

TABLE OF CONTENTS

[Predictive Analysis of Customer Spending Power (Customer Segmentation) 4](#_Toc15821501)

[Strategic Highlights / Business Case 4](#_Toc15821502)

[Objective 4](#_Toc15821503)

[Data Collection 5](#_Toc15821504)

[Data Preparation 5](#_Toc15821505)

[Dataset Details: 5](#_Toc15821506)

[Data Exploration & Visualization 6](#_Toc15821507)

[Basic statistical details of the dataset using describe() method: 6](#_Toc15821508)

[Histogram Plots: 6](#_Toc15821509)

[Boxplot & Swarmplots 7](#_Toc15821510)

[Summary observation on original dataset: 8](#_Toc15821511)

[Data Manipulation 8](#_Toc15821512)

[OLS Regression on initial dataset: 9](#_Toc15821513)

[Age Grouping 10](#_Toc15821514)

[Model Building 11](#_Toc15821515)

[Model I: OLS with Forward Selection 11](#_Toc15821516)

[Predicted Spending Score VS Actual Spending Score 11](#_Toc15821517)

[QQPlots for Residuals 12](#_Toc15821518)

[Model I: Results of regression analysis (OLS with Forward Selection): 12](#_Toc15821519)

[Coefficients (Slope) of Age and Annual Income explained: 13](#_Toc15821520)

[P-Value Explained: 13](#_Toc15821521)

[R-Squared Explained: 14](#_Toc15821522)

[Concluding Observations: 15](#_Toc15821523)

[Model II: Bayesian Linear Regression with PyMC3 15](#_Toc15821524)

[Create Model in PyMC3 and Sample from Posterior 17](#_Toc15821525)

[Examining Bayesian Linear Regression Results with Posterior Distributions 17](#_Toc15821526)

[Summary of MCMC Model for Different Age Groups: 23](#_Toc15821527)

[Interpretations 24](#_Toc15821528)

[Prediction of Response Variable: 24](#_Toc15821529)

[Prediction for Single Point: 26](#_Toc15821530)

[Conclusions 28](#_Toc15821531)

[Comparison of the two models’ Results: 28](#_Toc15821532)

[References 29](#_Toc15821533)

# Predictive Analysis of Customer Spending Power (Customer Segmentation)

## Strategic Highlights / Business Case

*To predict the spending power of the target customers by understanding the trends in the dataset and predicting spending score.*

**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.

|  |
| --- |
| ***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.*** |

## 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. Concluding on target customers from the customer base available in the 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)

## 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.

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

|  |  |
| --- | --- |
| 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 dataset using describe() method:

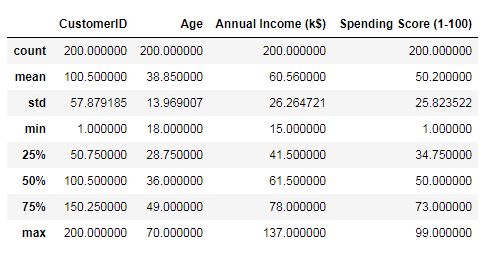


Figure 2: Basic Statistical Details of Dataset

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:

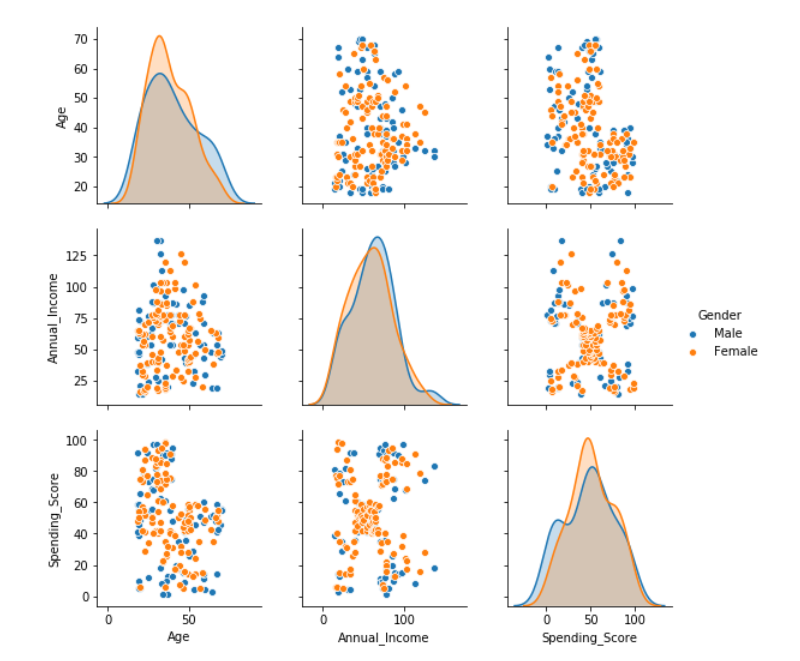
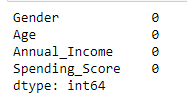


Figure 5: Pairplots on the Original Dataset

We see that Age and Annual Income are not collinear. There is no direct correlation that can be seen between the spending score and age, spending score and annual income. Thus, it would be appropriate to do forward selection to find the best combination of independent variables that yield the best results.

## Data Manipulation

Using the build in function, isnull(), we verified that there are no 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 be transformed before it can be used in data analysis. Therefore, we generated a new attribute, GenderGroup, by mapping 0 for female customers and 1 for male customers.

We then ran OLS regression to check which attributes can be dropped. This is done to further refine the list of attributes in order to make the best model.

### OLS Regression on initial dataset:

**Running OLS Regression on Annual Income, Age and Gender excluding intercept**

**Using the Formula**: Spending\_Score ~ Annual\_Income + Age + Gender-1

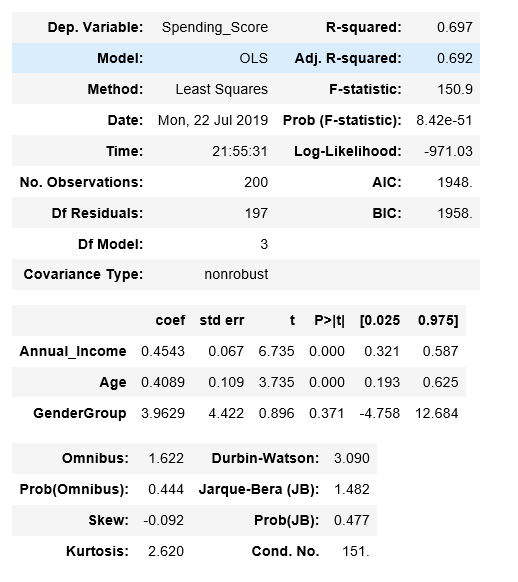


Figure 6: OLS Regression Results on Original Dataset without Intercept

**Interpretations:**

1. No clear linear relationship can be seen between the spending score and independent variables, 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. p-value for Gender Group is high (>0.05) showing that changes in predictor variable (Gender Group) are not associated with changes in the response variable (spending score). Since, Gender Group is not a significant factor for the change in spending score, we will rerun the model after dropping GenderGroup from our model.
4. Intercept can be excluded as no annual income suggests no spending power/score.

#### Excluding Intercept

We will exclude the intercept from our models, as we can safely say that at 0 on the x-axis for the predictor variables (Annual Income, Age and Gender) the response variable on Y-axis i.e. spending score will be 0.

### Age Grouping

We will divide the data into age groups and try to run regression analysis for each age group. This is done this 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

## Model I: OLS with Forward Selection

**Forward Selection**: Forward selection is a type of stepwise regression which begins with an empty model and adds in variables one by one. In each forward step, we add the one variable that gives the single best improvement to your 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- defined 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 described 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.

### Predicted Spending Score VS Actual Spending Score

A screenshot of a computer

Description automatically generated

Figure 8: Predicted vs Actual Spending Score

### QQPlots for Residuals

A close up of a map

Description automatically generated

Figure 9: 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.

## Model I: 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 | | | 89% |
| T-Value for Age: 0 | | T-Value for Annual Income: 0 |
| 26 – 35 | Spending\_Score = 1.6762 \* Age + 0.1718 \* Annual\_Income | | | 86% |
| T-Value for Age: 0 | T-Value for Annual Income: 0.11 | |
| 36 – 45 | Spending\_Score = 0.5828 \* Age + 0.2917 \* Annual\_Income | | | 70% |
| T-Value for Age: 0.13 | T-Value for Annual Income: 0.15 | |
| 46 – 55 | Spending\_Score = 0.7401 \* Age | | | 83% |
| T-Value for Age: 0 | | |
| 56 – 65 | Spending\_Score = 0.5358 \* Age | | | 71% |
| T-Value for Age: 0 | | |
| Above 66 | Spending\_Score = 0.2173 \* Age + 0.6344 \* Annual\_Income | | | 97% |
| T-Value for Age: 0.01 | T-Value for Annual Income: 0.19 | |

### Coefficients (Slope) of Age and Annual Income explained:

1. For Age Group Below 25:

a. 1 unit increase in Age results in 4.0296 unit increase in Spending Score.

b. 1 unit increase in Annual Income results in 0.6769 unit decrease in Spending Score.

1. For Age Group 26 - 35:

a. 1 unit increase in Age results in 1.6762 unit increase in Spending Score.

b. 1 unit increase in Annual Income results in 0.1718 unit increase in Spending Score.

1. For Age Group 36 - 45:

a. 1 unit increase in Age results in 0.5828 unit increase in Spending Score.

b. 1 unit increase in Annual Income results in 0.2917 unit increase in Spending Score.

1. For Age Group 46 - 55:

a. 1 unit increase in Age results in 0.7401 unit increase in Spending Score.

b. Based on Forward Selection results Annual Income is not used in the model for Age Group 46 - 55.

1. For Age Group 56 - 65:

a. 1 unit increase in Age results in 0.5358 unit increase in Spending Score.

b. Based on Forward Selection results Annual Income is not used in the model for this age group.

1. For Age Group Above 65: a. 1 unit increase in Age results in 0.2173 unit increase in Spending Score. b. 1 unit increase in Annual Income results in 0.6344 unit increase in Spending Score.

### P-Value Explained:

1. For Age Group Below 25:
   1. P-Value for Age is 0.0 (<0.05 – Age is a statistically significantly parameter in this case)
   2. P-Value for Annual Income is 0.0 (<0.05 – Annual Income is a statistically significantly parameter in this case)
2. For Age Group 26 - 35:
   1. P-Value for Age is 0.00 (<0.05 – Age is a statistically significantly parameter in this case)
   2. P-Value for Annual Income is 0.11 (>0.05 – Annual Income is not a statistically significantly parameter in this case)
3. For Age Group 36 - 45:
   1. P-Value for Age is 0.13 (<0.05 – Age is not a statistically significantly parameter in this case)
   2. P-Value for Annual Income is 0.15 (>0.05 – Annual Income is not a statistically significantly parameter in this case)
4. For Age Group 46 - 55:
   1. P-Value for Age is 0.00 (<0.05 – Age is a statistically significantly parameter in this case)
   2. Annual Income is not considered in the model.
5. For Age Group 56 - 65:
   1. P-Value for Age is 0.00 (<0.05 – Age is a statistically significantly parameter in this case)
   2. P-Value for Annual Income is 0.01 (<0.05 – Annual Income is a statistically significantly parameter in this case)
6. For Age Group Above 65:
   1. P-Value for Age is 0.01 (<0.05 – Age is a statistically significantly parameter in this case)
   2. P-Value for Annual Income is 0.19 (>0.05 – Annual Income is not a statistically significantly parameter in this case)

### R-Squared Explained:

1. For Age Group Below 25: R-squared value is 0.89. 89% variance in spending score can be explained by the regression model using Age and Annual Income as the dependent variables.
2. For Age Group 26 - 35: R-squared value is 0.86. 86% variance in spending score can be explained by the regression model using Age and Annual Income as the dependent variables.
3. For Age Group 36 - 45: R-squared value is 0.7. 70% variance in spending score can be explained by the regression model using Age and Annual Income as the dependent variables.
4. For Age Group 46 - 55: R-squared value is 0.83. 83% variance in spending score can be explained by the regression model using Age as the dependent variables.
5. For Age Group 56 - 65: R-squared value is 0.71. 71% variance in spending score can be explained by the regression model using Age as the dependent variables.
6. For Age Group Above 65: R-squared value is 0.97. 97% variance in spending score can be explained by the regression model using Age and Annual Income as the dependent variables.

### Concluding 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 the included variables, from the forward selection, generated the best predicting for that particular age group. Any other combination would have yielded an even lower 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 years old.

6. There are not many data points 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.

## Model 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.

We do this primarily because:

1. We have a limited dataset
2. Some facts may be more likely than others, but that information may not be contained in the data we are using for modeling.
3. We are interested in knowing how likely certain facts are

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 characterizes 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.

Critical Advantages of Bayesian estimation:

* + Priors: We can quantify any prior knowledge we might have by placing priors on the parameters.
  + Quantifying uncertainty: We do not get a single estimate of β but instead a complete posterior distribution about how likely different values of β are.

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

A screenshot of a cell phone

Description automatically generated

Figure 10: Bayesian Regression - Heatmap

Looking at the heatmap we can see Age, Annual Income and GenderGroup show statistical relevance to Spending score as these show up with darker shades as shown by the scale. We see a strong correlation between Age and Spending Score, and some correlation between Annual Income and Spending Score.

### Create Model in PyMC3 and Sample from Posterior

We now build the model using the formula 'Spending\_Score ~ Age + Annual\_Income' as evident from heat map, and a normal distribution for the data likelihood. Then, we let a Markov Chain Monte Carlo algorithm draw samples from the posterior to approximate the posterior for each of the model parameters.

* The sampling algorithm chosen by PYMC3 module is NUTS sampler i.e. No U-turn Sampler.

We draw 20000 sample sets with 2 chains and run the GLM model for the formula: Spending\_Score ~ Age + Annual\_Income -1

As seen from the formula, we use Age and Annual Income and remove the intercept from consideration.

We do this for the same 6 age groups as identified before (below 25, 26 – 35, 36 – 45, 46 – 55, 56 – 65, Above 65).

### Examining Bayesian Linear Regression Results with Posterior Distributions

#### Age Group: Below 25

A close up of a mans face

Description automatically generated

Figure 11: Linear Trace - Below 25

A close up of a map

Description automatically generated

Figure 12: Posterior - Below 25

#### Age Group: 26 – 35

A close up of a device

Description automatically generated

Figure 13: Linear Trace - 26 to 35

A close up of a map

Description automatically generated

Figure 14: Posterior - 26 - 35

#### Age Group: 36 – 45

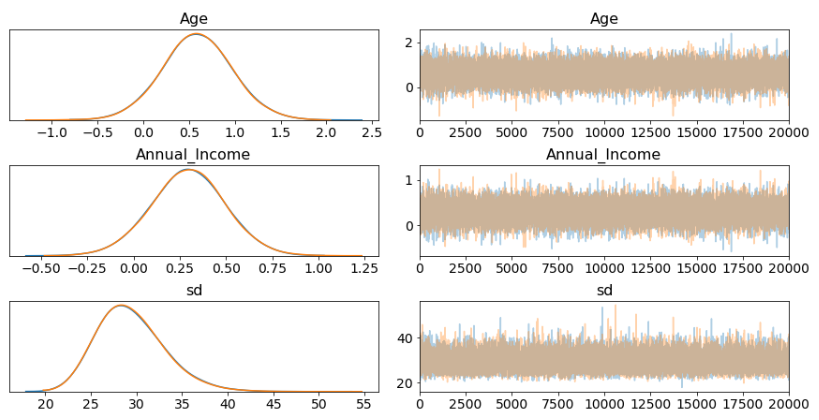


Figure 15: Linear Trace - 36 to 45

A close up of a map

Description automatically generated

Figure 16: Posterior - 36 - 45

#### Age Group: 46 – 55

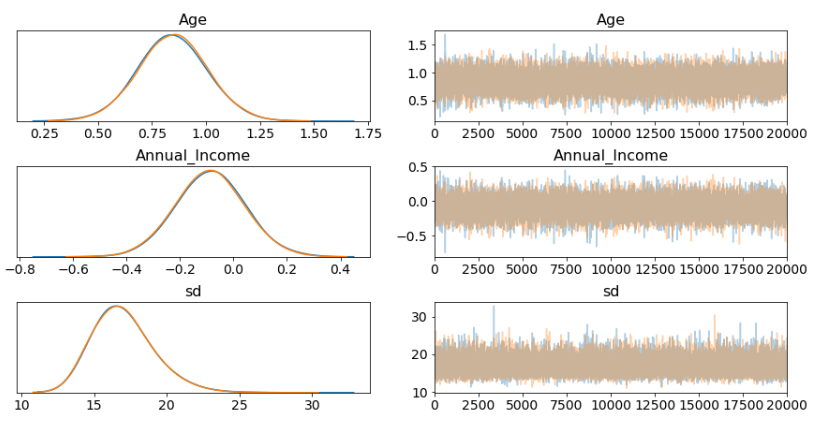


Figure 17: Linear Trace - 46 to 55

A close up of a map

Description automatically generated

Figure 18: Posterior - 46 - 55

#### Age Group: 56 – 65

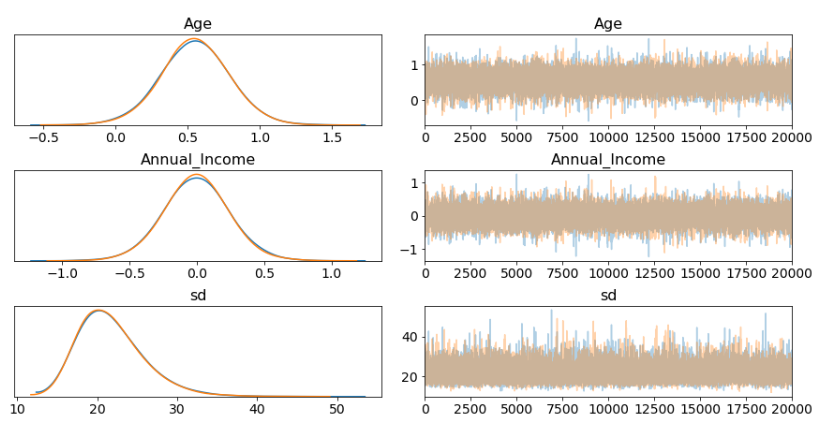


Figure 19: Linear Trace - 56 to 65

A close up of text on a white background

Description automatically generated

Figure 20: Posterior - 56 - 65

#### Age Group: Above 65

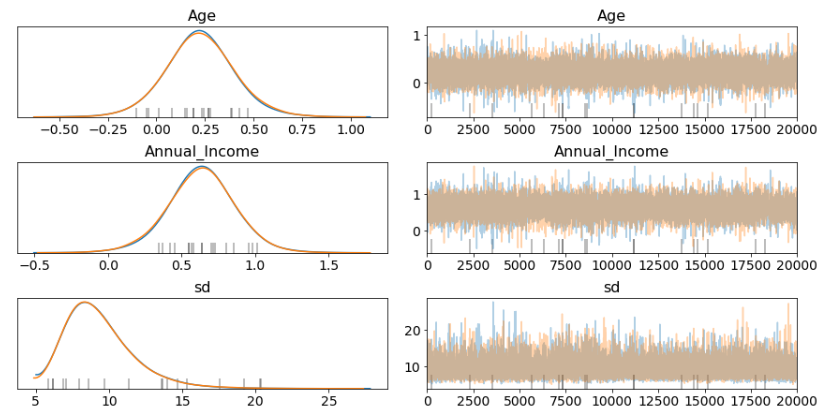


Figure 21: Linear Trace - Above 65

A close up of text on a white background

Description automatically generated

Figure 22: Posterior - Above 65

Bayesian inference does not give us only one best fitting line (as maximum likelihood does) but rather a whole posterior distribution of likely parameters.

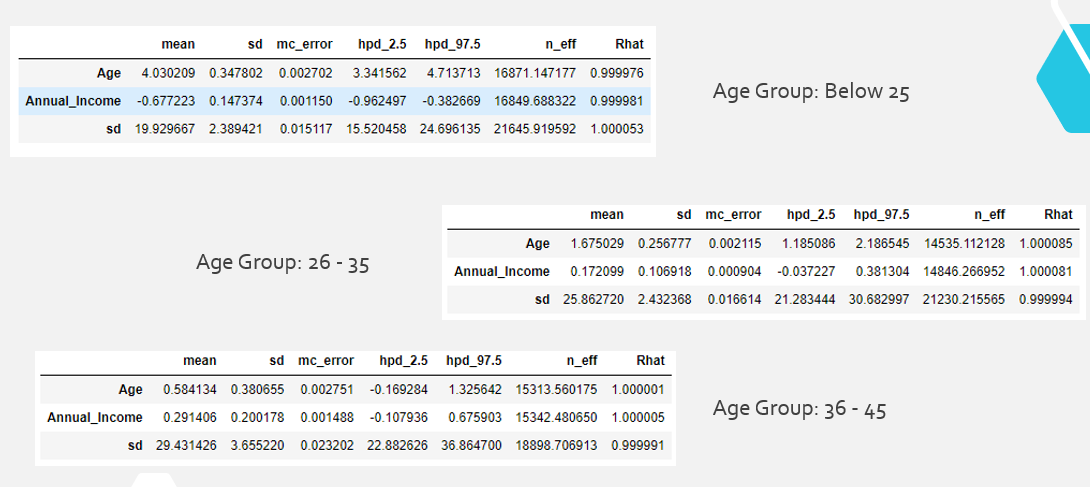
**Linear Trace Interpretation**: The left side of the traceplot shows the marginal posterior:

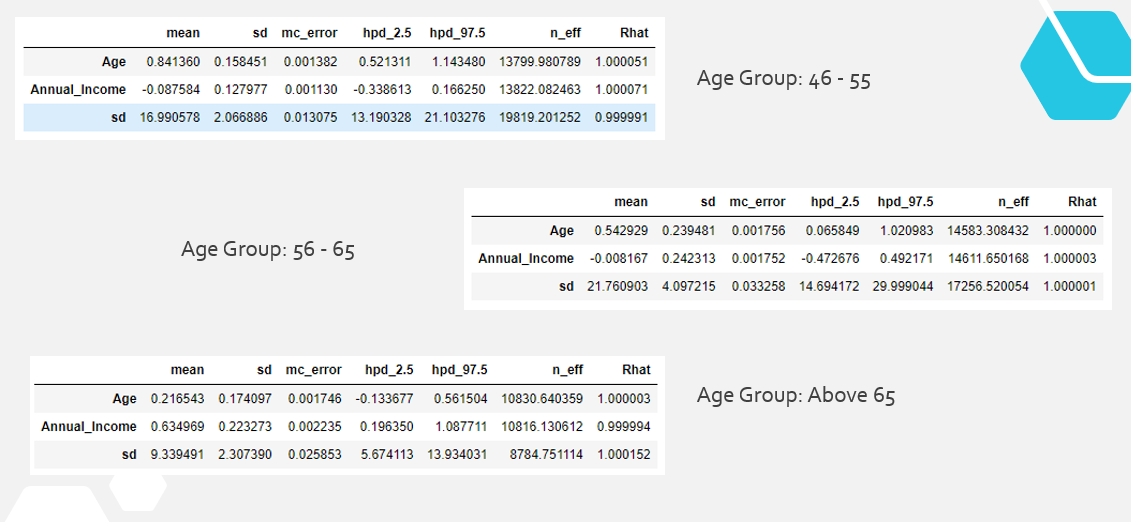
* X-Axis: The values for the variable
* Y-Axis: The probability for the variable (as determined by sampling)
* The different colored lines indicate that we performed two chains of Markov Chain Monte Carlo.
* From the left side we can see that there is a range of values for each weight.
* The right side shows the different sample values drawn as the sampling process runs.

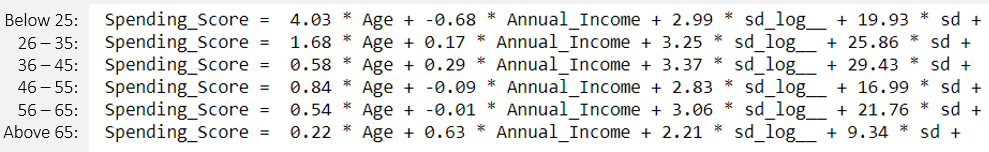
There are a couple of things to see here:

* Our sampling chains for the individual parameters (left side) seem well converged and stationary (there are no large drifts or other odd patterns).
* The maximum posterior estimate of each variable (the peak in the left side distributions) is very close to the true parameters used to generate the data (x is the regression coefficient and sigma is the standard deviation of our normal).

### Summary of MCMC Model for Different Age Groups:







### Interpretations

Based on the sign and location of the weights, we can make the following inferences regarding the features in our dataset:

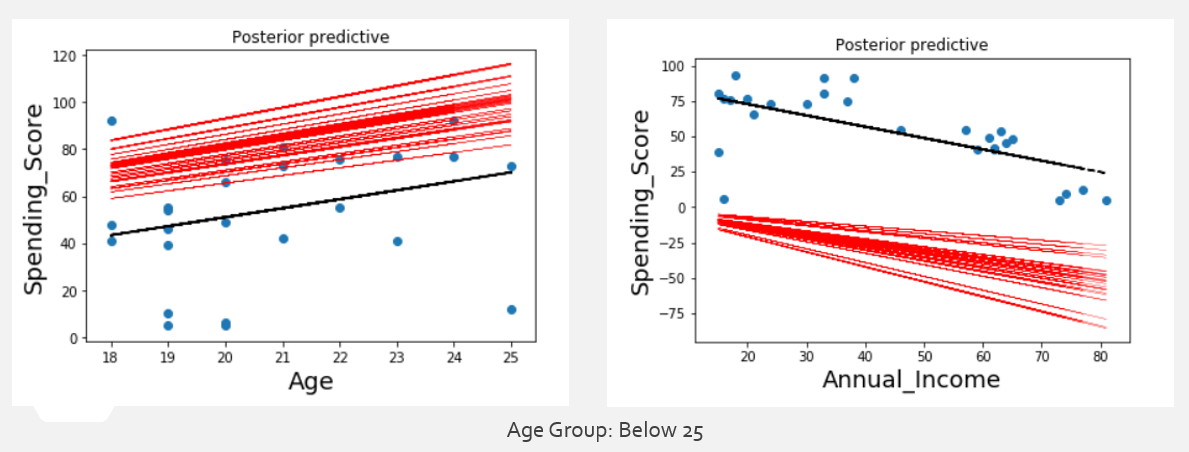
* Based on Agegroup, Below 25, Age seems to have positive relation with spending score while Annual income has negative impact.
* In Agegroup between 26-45, both Age and Annual income has postive impact on spending score.
* When we go further down the age group with ages 56-65, we see a positive impact of age on spending score while annual income is negative.
* In Agegroup, above 66, both the factors has positive impact on age.

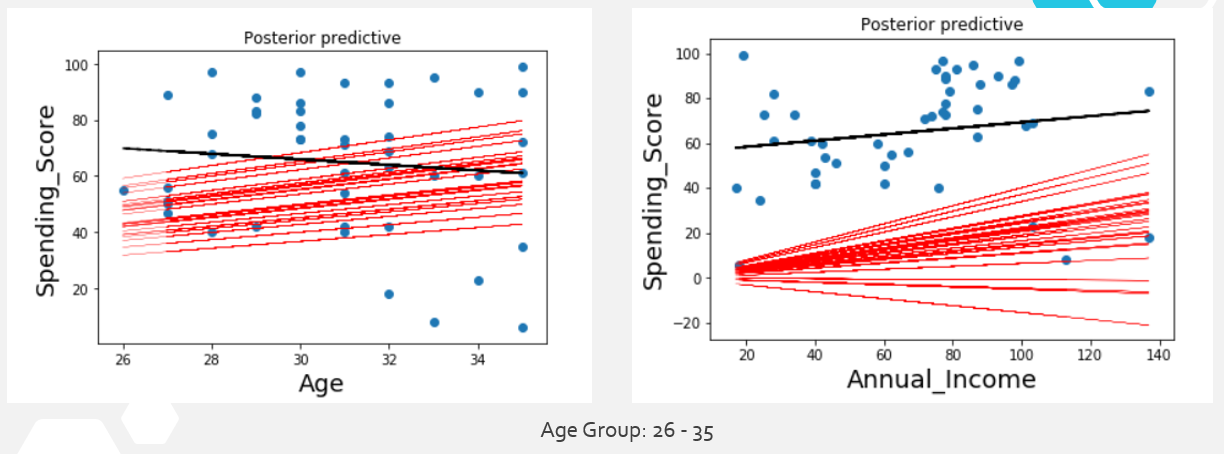
### Prediction of Response Variable:

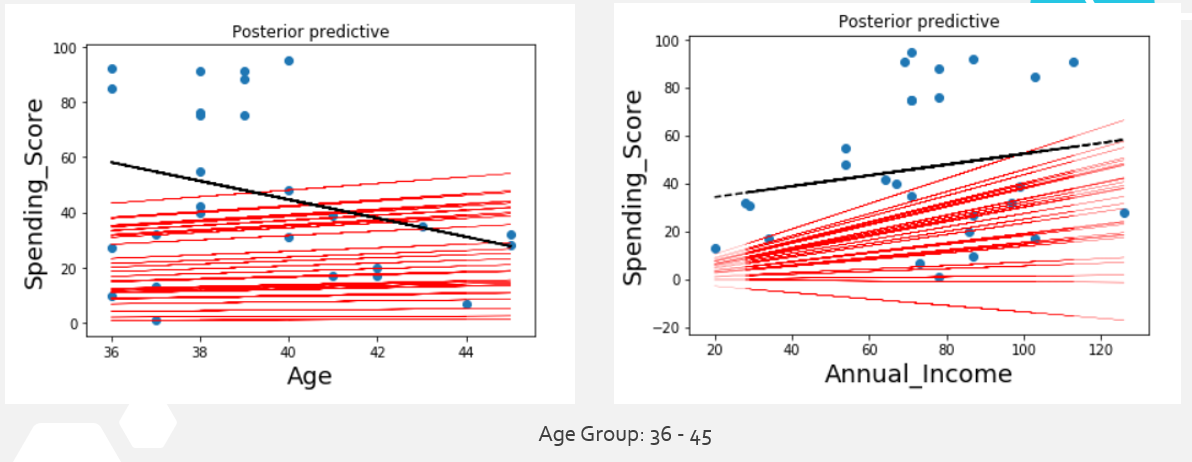
We create the posterior plots which provide insight into the sample’s posterior predictive regression lines.

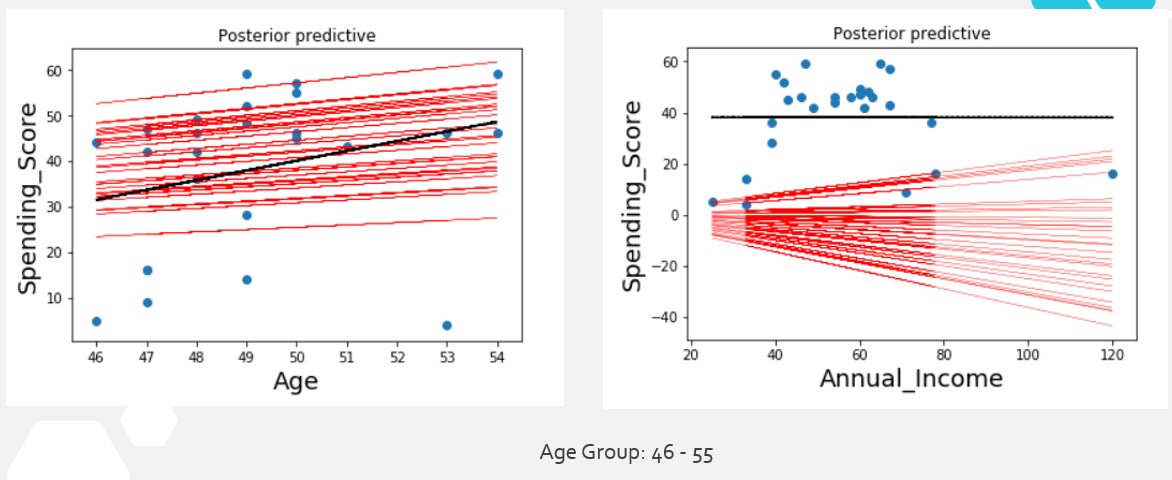
In the GLM we do not only have one best fitting regression line, but many. A posterior predictive plot takes multiple samples from the posterior (intercepts and slopes) and plots a regression line for each of them.

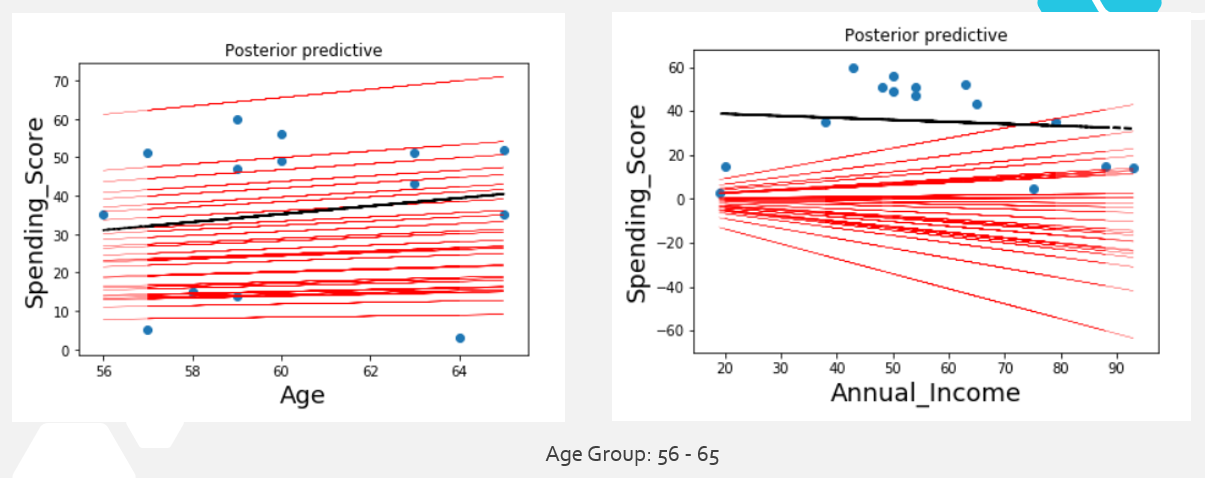
Estimated regression lines are very similar to the true regression line. But since we only have limited data, we have uncertainty in our estimates, here expressed by the variability of the lines.

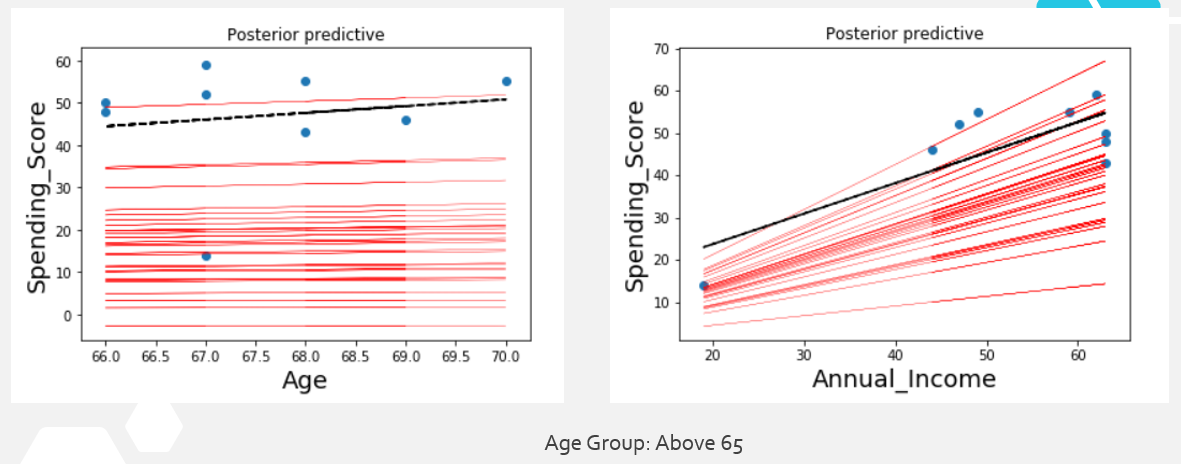






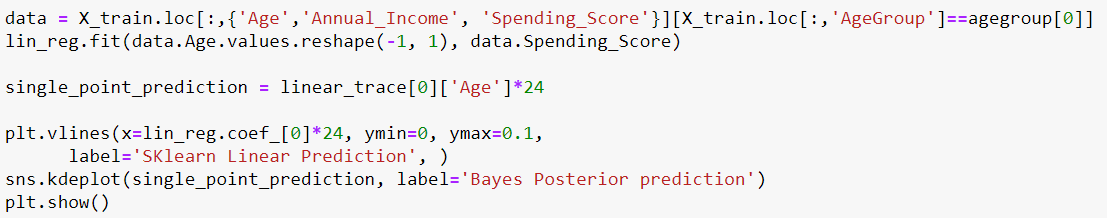






### Prediction for Single Point:

Below is the point prediction (for Age = 24, using Bayesian Regression. We see that Bayesian prediction mean value almost coincides with the linear regression output. The data should be rich enough to predict correctly.



A close up of a map

Description automatically generated

Figure 23: Bayesian Regression – Single Point Prediction

# Conclusions

## Comparison of the two models’ Results:

### OLS Regression Results:

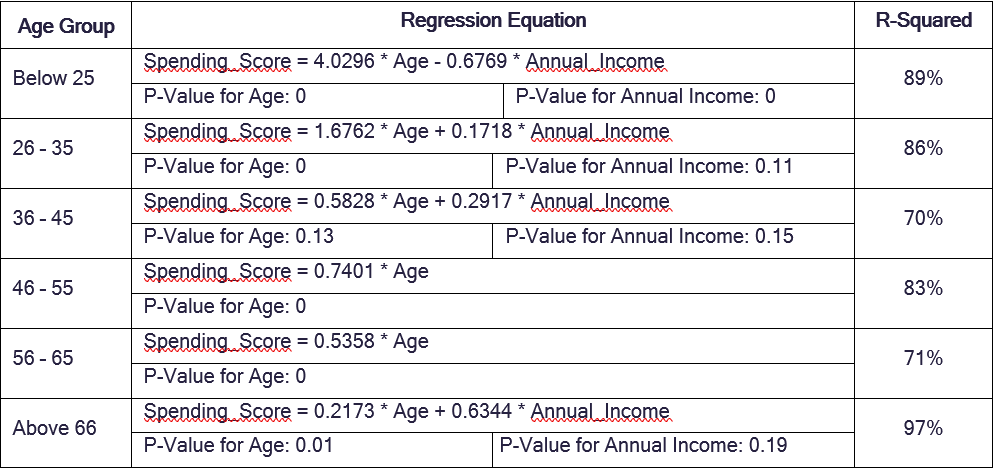


Figure 24: OLS Regression Resulting Equations

### Bayesian Regression Results:

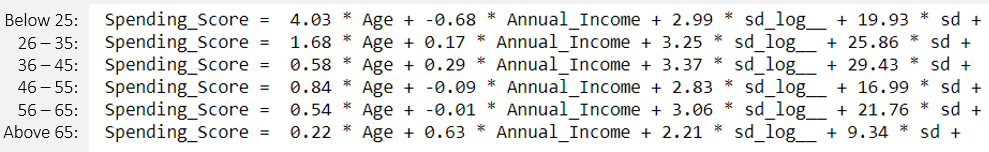


Figure 25: Bayesian Regression Resulting Equations

As we can see from above, coefficients from both regression models are very close. Below is a summary of comparison:

* Below 25 Age Group: Coefficient for Age from OLS is 4.0296, which is very close to Age coefficient of 4.03 from Bayesian Regression.

Coefficient for Annual Income from OLS is (-0.6769), which is very close to the -0.6769 value from Bayesian Regression.

Overall, shows a very similar dependency on Age and Annual Income in this group.

* 26 – 35 Age Group: Coefficient for Age from OLS is 1.6762, which is very close to Age coefficient of 1.68 from Bayesian Regression.

Coefficient for Annual Income from OLS is 0.1718, which is very close to the 0.17 value from Bayesian Regression.

Overall, shows a very similar dependency on Age and Annual Income in this group.

* 36 – 45 Age Group: Coefficient for Age from OLS is 0.5828, which is very close to Age coefficient of 0.58 from Bayesian Regression.

Coefficient for Annual Income from OLS is 0.2917, which is very close to the 0.29 value from Bayesian Regression.

Overall, shows a very similar dependency on Age and Annual Income in this group.

* 46 – 55 Age Group: Coefficient for Age from OLS is 0.7401, which is only different from Age coefficient of 0.84 from Bayesian Regression by 0.09 units.

OLS regression for this age group does not use Annual Income as one of the independent variables, determined by forward selection. However, in Bayesian Regression we have used both age and annual income. Coefficient for Annual Income is -0.09 from Bayesian Regression, indicating a very minimal and a negative dependency of spending score on annual income.

Overall both models show very similar results.

* 56 – 65 Age Group: Coefficient for Age from OLS is 0.5358, which is very close to Age coefficient of 0.54 from Bayesian Regression.

OLS regression for this age group does not use Annual Income as one of the independent variables, determined by forward selection. However, in Bayesian Regression we have used both age and annual income. Coefficient for Annual Income is -0.01 from Bayesian Regression, indicating a very minimal and a negative dependency of spending score on annual income.

Overall, both models show very similar results.

* Above 65 Age Group: Coefficient for Age from OLS is 0.2173, which is very close to Age coefficient of 0.22 from Bayesian Regression.

Coefficient for Annual Income from OLS is 0.6344, which is very close to the 0.63 value from Bayesian Regression.

Overall, shows a very similar dependency on Age and Annual Income in this group.

# References

* <https://searchcustomerexperience.techtarget.com/definition/customer-segmentation>
* <https://towardsdatascience.com/markov-chain-monte-carlo-in-python-44f7e609be98>
* <https://www.quantstart.com/articles/Bayesian-Linear-Regression-Models-with-PyMC3>
* <https://www.quantstart.com/articles/Markov-Chain-Monte-Carlo-for-Bayesian-Inference-The-Metropolis-Algorithm>
* <https://towardsdatascience.com/bayesian-linear-regression-in-python-using-machine-learning-to-predict-student-grades-part-2-b72059a8ac7e>
* <http://people.duke.edu/~ccc14/sta-663-2016/16C_PyMC3.html>
* <https://people.duke.edu/~ccc14/sta-663/PyMC3.html>