Customer segmentation with k-means clustering

Nathan Wong

June 2019

INTRODUCTION

This dataset contains information about mall customers who have membership cards. The purpose of this project is to identify groups of customers based on their gender, age, annual income, and spending score (customer behavior and purchasing data). Purchasing patterns can be examined by categorizing customers. In order to maximize sales, different types of advertisements, discounts, and rewards can target distinct groups. For example, a company can use affinity analysis to recommend products to a customer based on the customer's purchasing history and the purchasing history of other customers who bought the same item.

More information can be found here:

https://www.kaggle.com/vjchoudhary7/customer-segmentation-tutorial-in-python

METHODS

Dataset

This dataset has 200 records of customer data. Each record includes the customer's gender, age, annual income, and spending score. The dataset does not have any missing values.

Demographics

There are 112 females and 88 males (see Figure 1). The mean age is approximately 39 years old, and the median age is 36 years old. The age range is 18 years old to 70 years old (see Figure 2). The mean annual income score is 60.56, and the median is 61.50 points. The lowest annual income score is 15 points, and the highest is 137 points (see Figure 3). The mean spending score is 50.20 points, and the median is 50 points. The lowest spending score is 1 point, and the highest is 99 points (see Figure 4).

Analysis

An independent samples t-test is used to compare the spending scores of females versus males. To ensure the independent samples t-test assumptions are met, Bartlett's test is used to compare the variance of the female group and the variance of the male group. Bartlett's test results indicate the two groups have equal variance ($p > .05$). The independent samples t-test results indicate there is no difference in spending score between females and males; $t(198)=0.82$, $p = 0.414$ (see Figure 5).

After plotting age and spending score, it appears that customers who are under 40 years old have a much higher spending score range than customers who are above 40 years old (see Figure 6). A multiple linear regression was calculated to predict spending score based on age and annual income. Results indicate there is a significant relationship between age and spending score (*p* = .000*)*; there is no relationship between annual income and spending score (*p* = .931).

Results from k-means clustering indicate 5 unique groups (see Table 1 and Figure 7).

Table 1: Annual Income vs. Spending Score

	Annual Income	**Spending Score**	**Color (see Figure 7)**
Group 1 (N =23)	Low	Low	Black
Group 2 (N = 23)	Low	High	Red
Group 3 (N = 36)	High	Low	Green
Group 4 (N = 39)	High	High	Blue
Group 5 (N = 79)	Medium	Medium	Light Blue

Figure 1: Gender

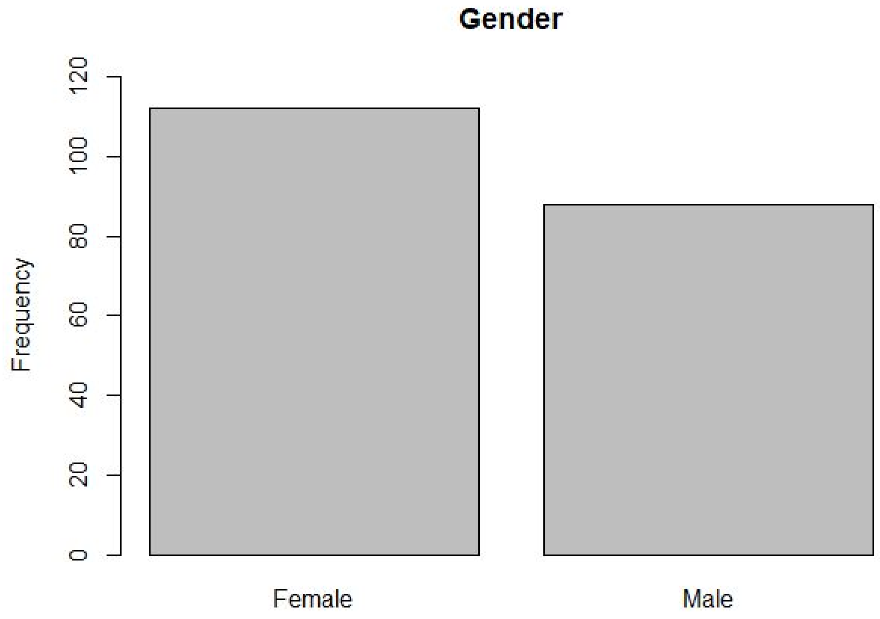


Figure 2: Age

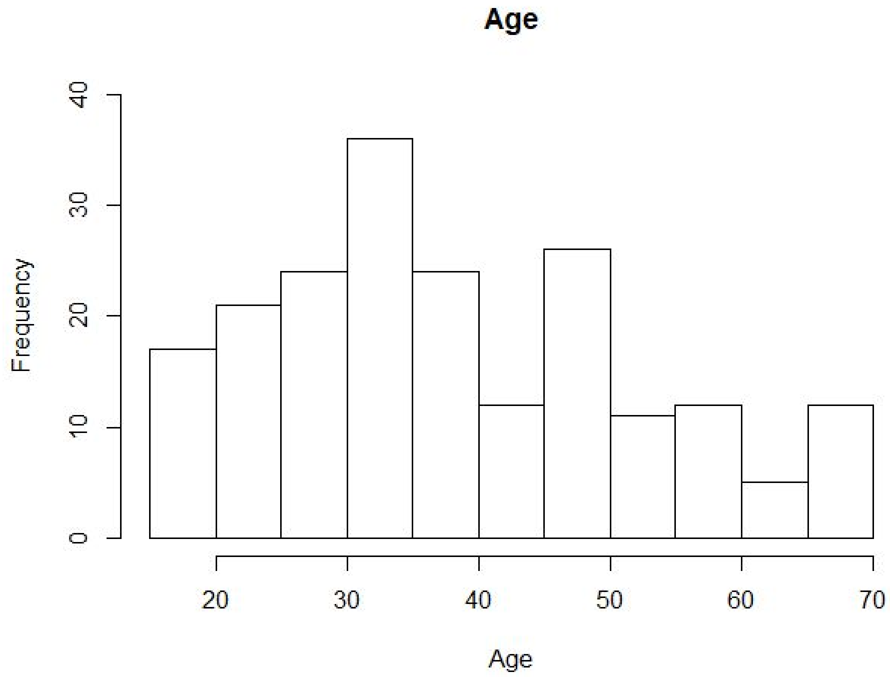


Figure 3: Annual Income

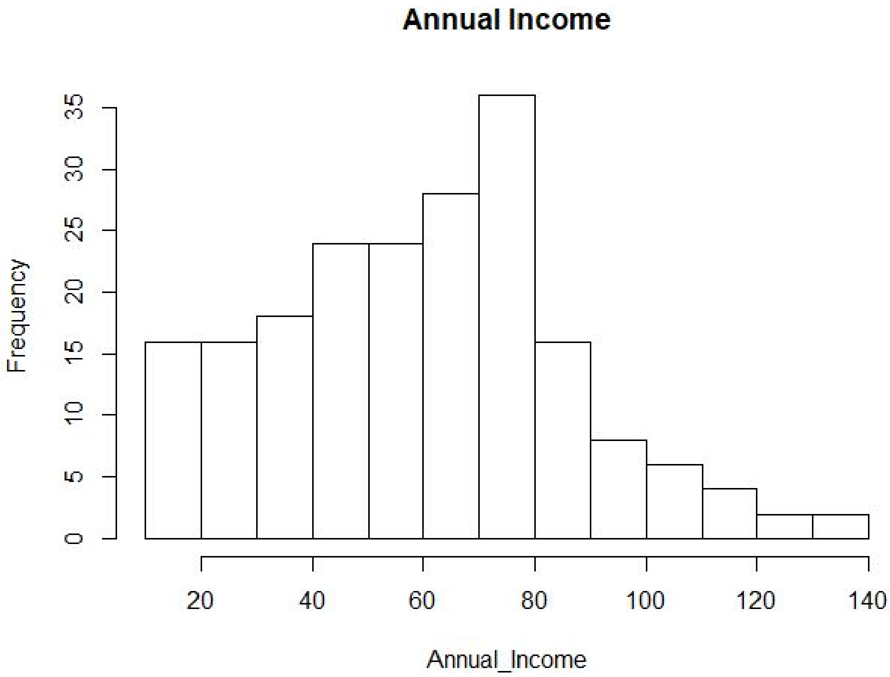


Figure 4: Spending Score

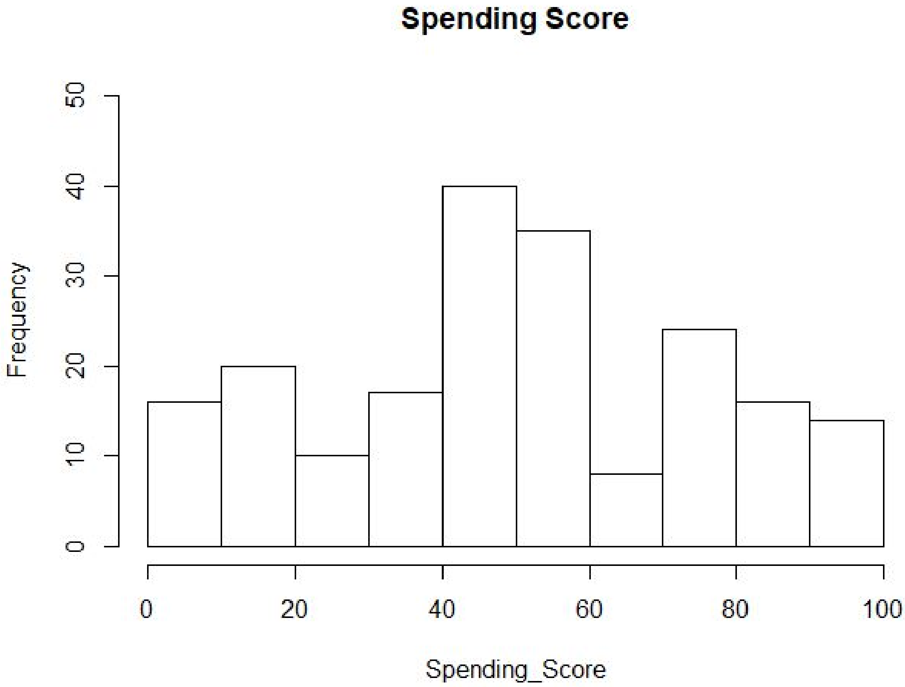


Figure 5: Gender vs. Spending Score

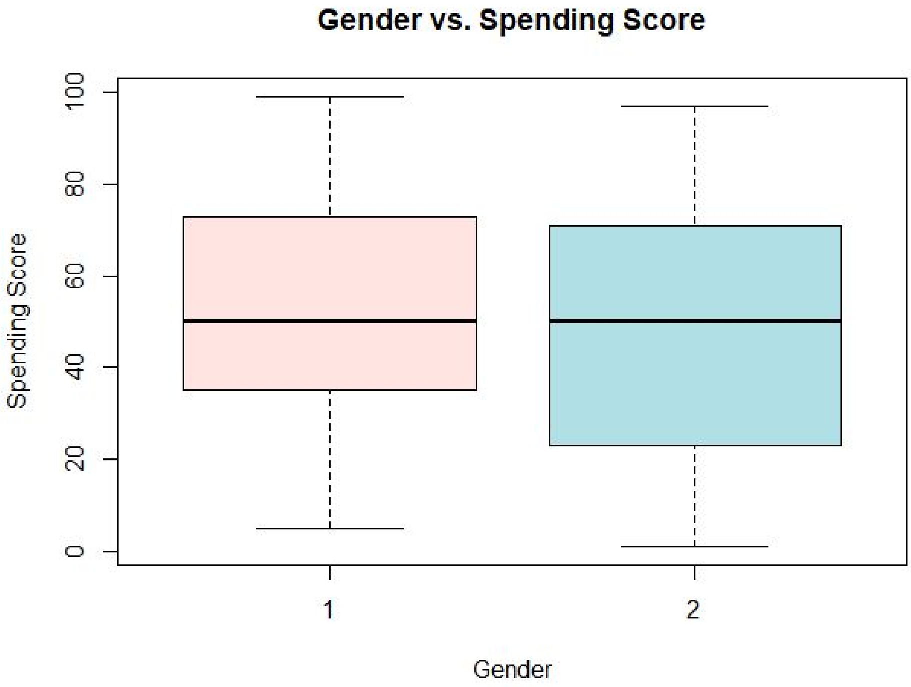


Figure 6: Age vs. Spending Score

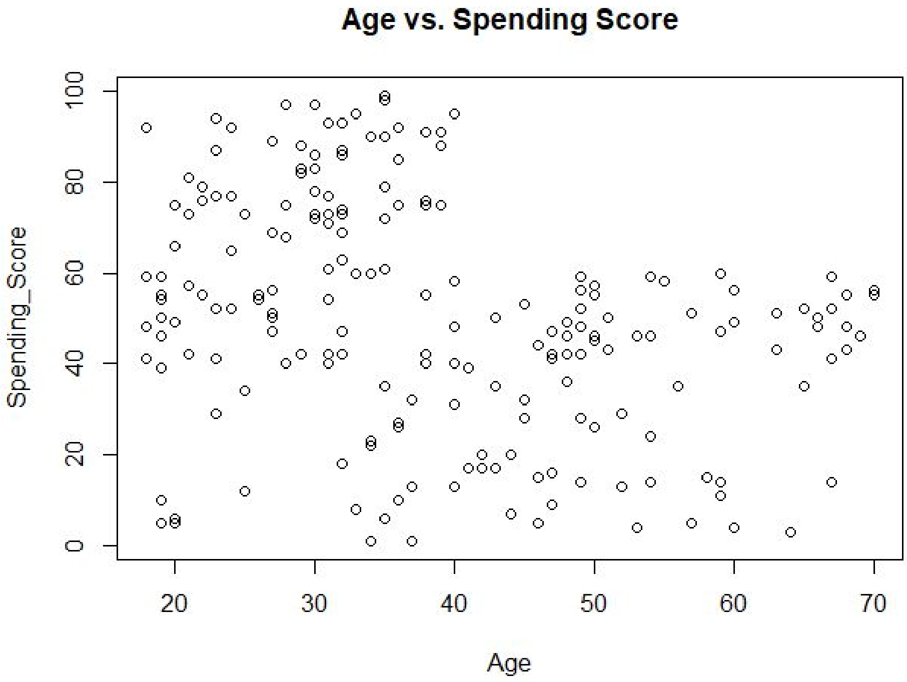
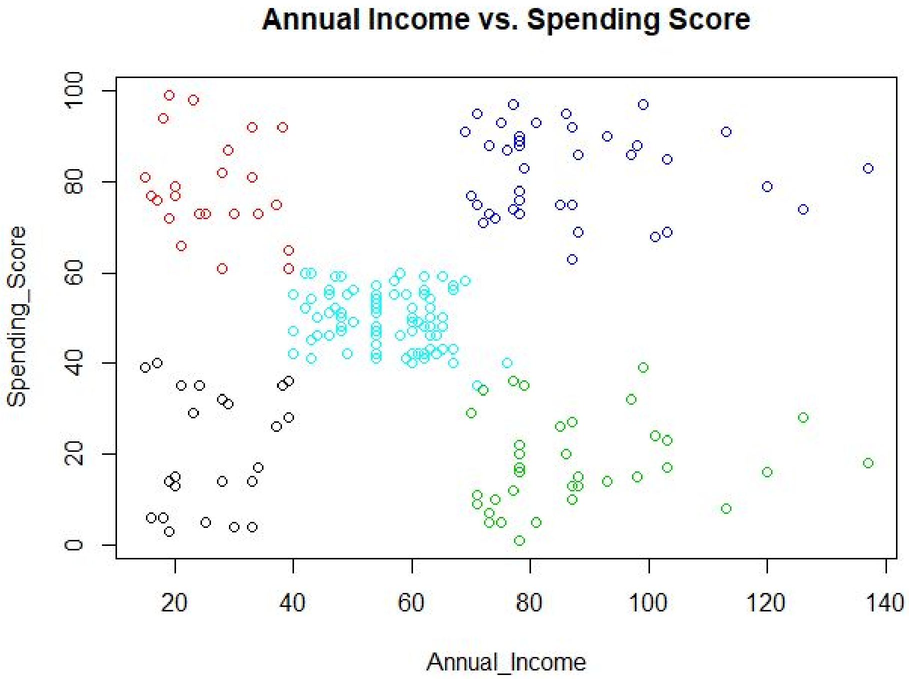


Figure 7: Annual Income vs. Spending Score


Annual Income vs. Spending Score
Spending_Score
Annual_Income