Consumer spending power analysis

會計二 110205090 李亦凱
經濟二 110510207 Yi-hui Li

報告目錄

01

Background

- 資料來源及出處
- 研究主題

02

Descriptive Statistics

- 離散變數敘述統計—性別、職業、家庭扶養數、工作經驗
- 連續變數敘述統計—所得、年齡、消費能力

03

Statistical Inference

- 性別、職業交叉消費能力
- 所得、年齡與消費能力的迴歸分析
- 綜合變數分析
- 預測

04

Conclusion

- 結論敘述

01
Background

- 資料來源及出處
- 研究主題

Sources and origins of the data

This dataset comes from the Kaggle dataset. It is a virtual dataset created based on customers of an ideal store. This data allows businesses to better understand customer spending habits, enabling them to make future price adjustments, modify in-store product types or formats, and other sales strategies.

Data Overview

CustomerID	Gender	Age	Income	SpendingScore	Profession	WorkExperience	FamilySize
1	Male	19	15000	39	Healthcare	1	4
2	Male	21	35000	81	Engineer	3	3
3	Female	20	86000	6	Engineer	1	1
4	Female	23	59000	77	Lawyer	0	2
5	Female	31	38000	40	Entertainment	2	6

(Source：https://www.kaggle.com/datasets/datascientistanna/customers-dataset?resource=download)

Research Topic

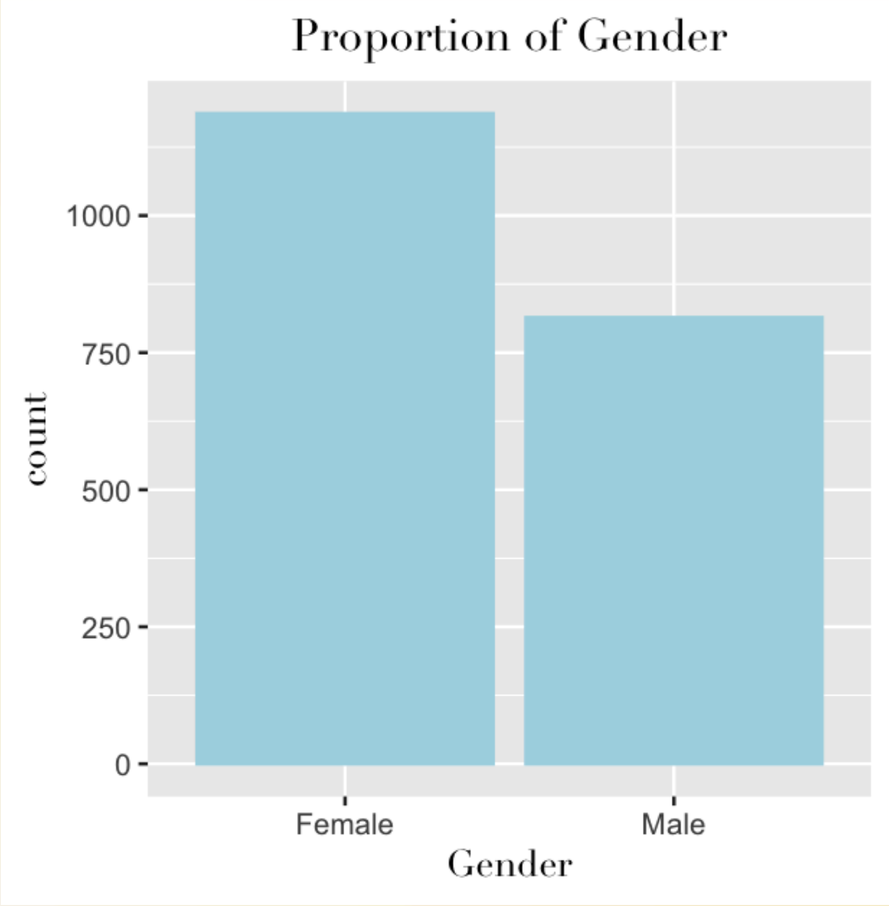
This research group explores the consumption power of different groups.

The main variable is the Spending Score. It involves cross-analysis of Spending Score with variables such as gender, age, occupation, annual income, or number of family members. The types of variables used in cross-analysis with spending score include binary (e.g., Gender), multivariate (e.g., Occupation), and continuous (e.g., Annual Income), which aligns well with the analytical techniques learned this semester, such as comparing two sets of continuous data, categorical data analysis, and simple regression analysis.

02
Descriptive Statistics

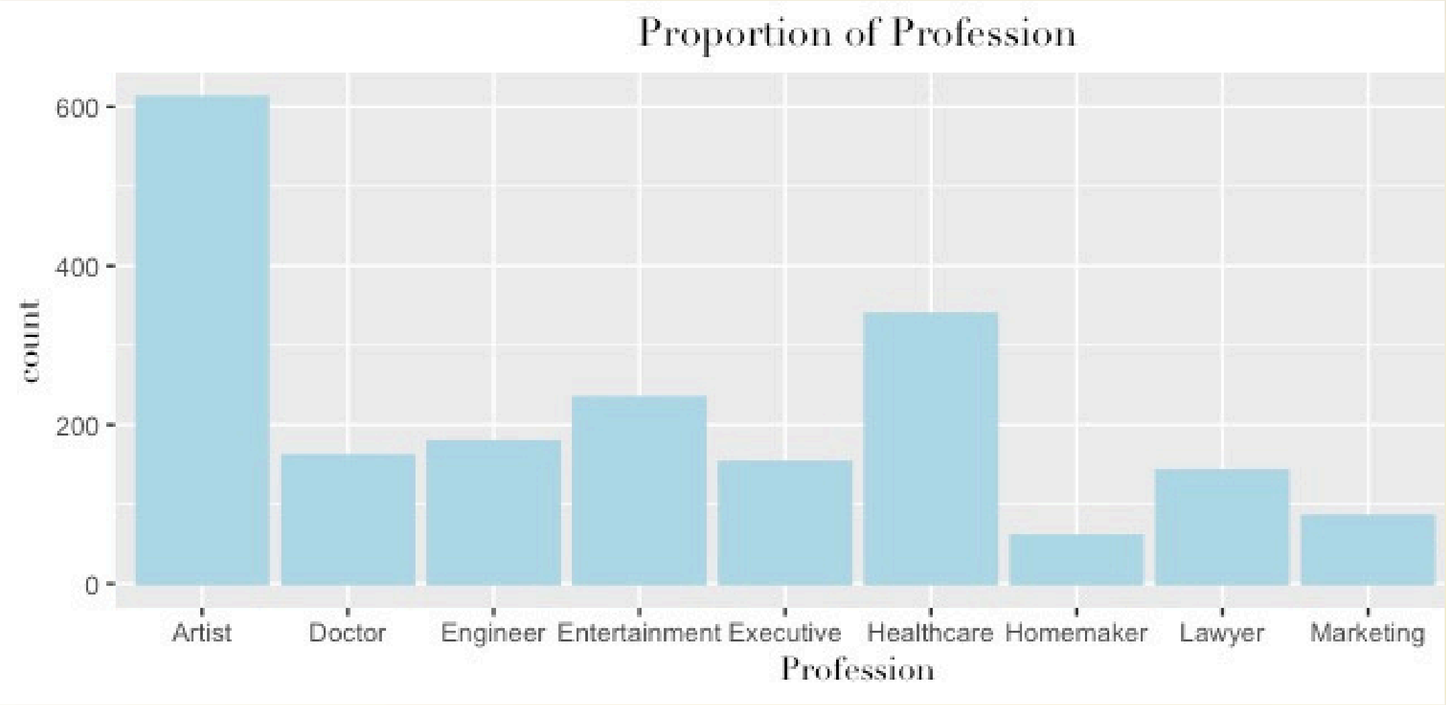
- 離散變數敘述統計—性別、職業、家庭扶養數、工作經驗
- 連續變數敘述統計—所得、年齡、消費能力

Discrete Data: Gender



This data sample contains 2,000 entries. 1,180 entries are from women, accounting for 59% of the total sample; while 820 entries are from men, accounting for 41%. It can be seen that women make up a larger proportion of this sample than men.

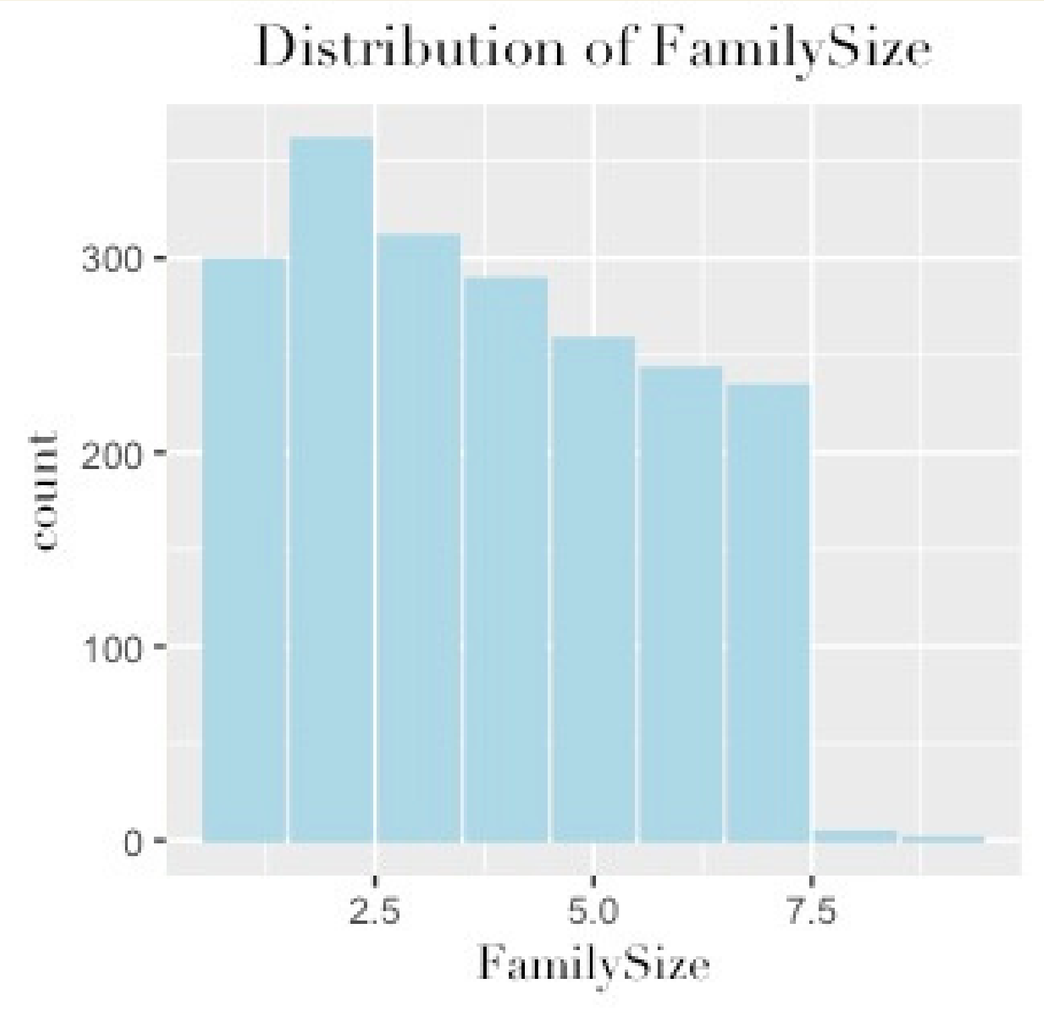
Discrete data: occupation



There are eight occupational categories in total: Artist, Healthcare, Executive, Engineer, Lawyer, Doctor, Homemaker, and Entertainment.

Artist has the highest percentage at 31%, followed by Healthcare at 17%.

Discrete data: Family Size

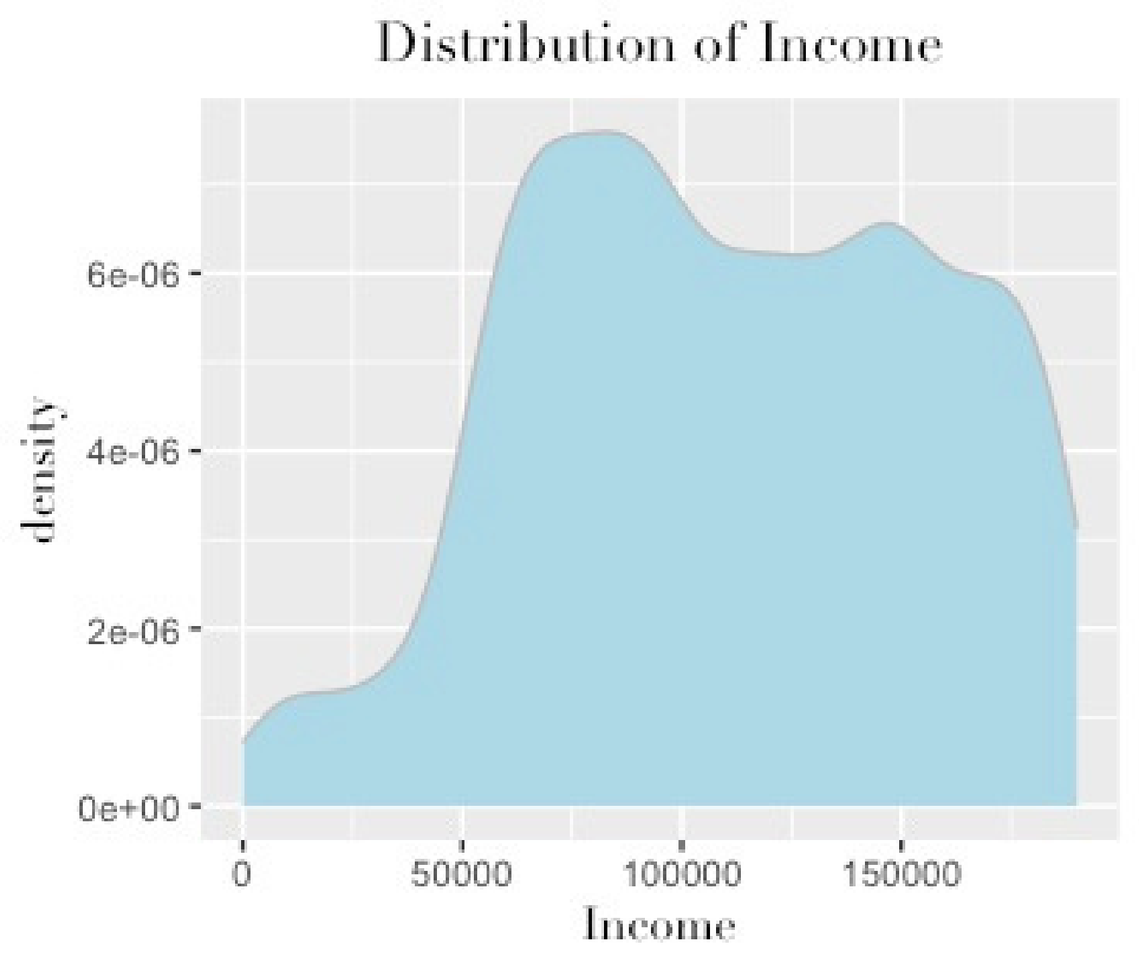


The highest number of cases (361 cases) had 2 dependents, accounting for about 18%; followed by cases with 3 dependents (311 cases) and cases with 1 dependent (299 cases).

Discrete Data: Working Experience

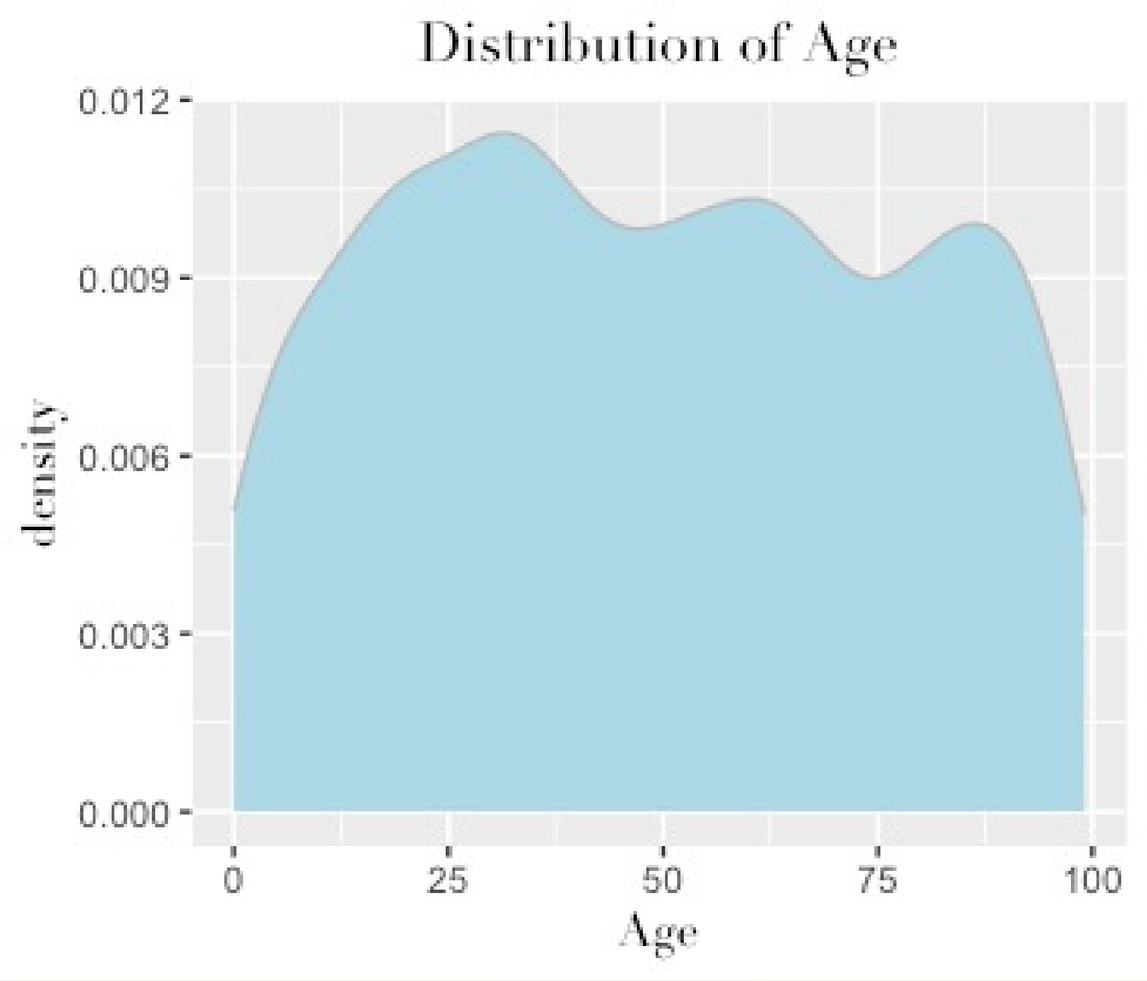
Those with one work experience and those with no work experience constituted the majority, with 470 and 431 records respectively, totaling approximately 45% of the total data; those with more than 10 work experiences were relatively few.

Continuous Data: Income



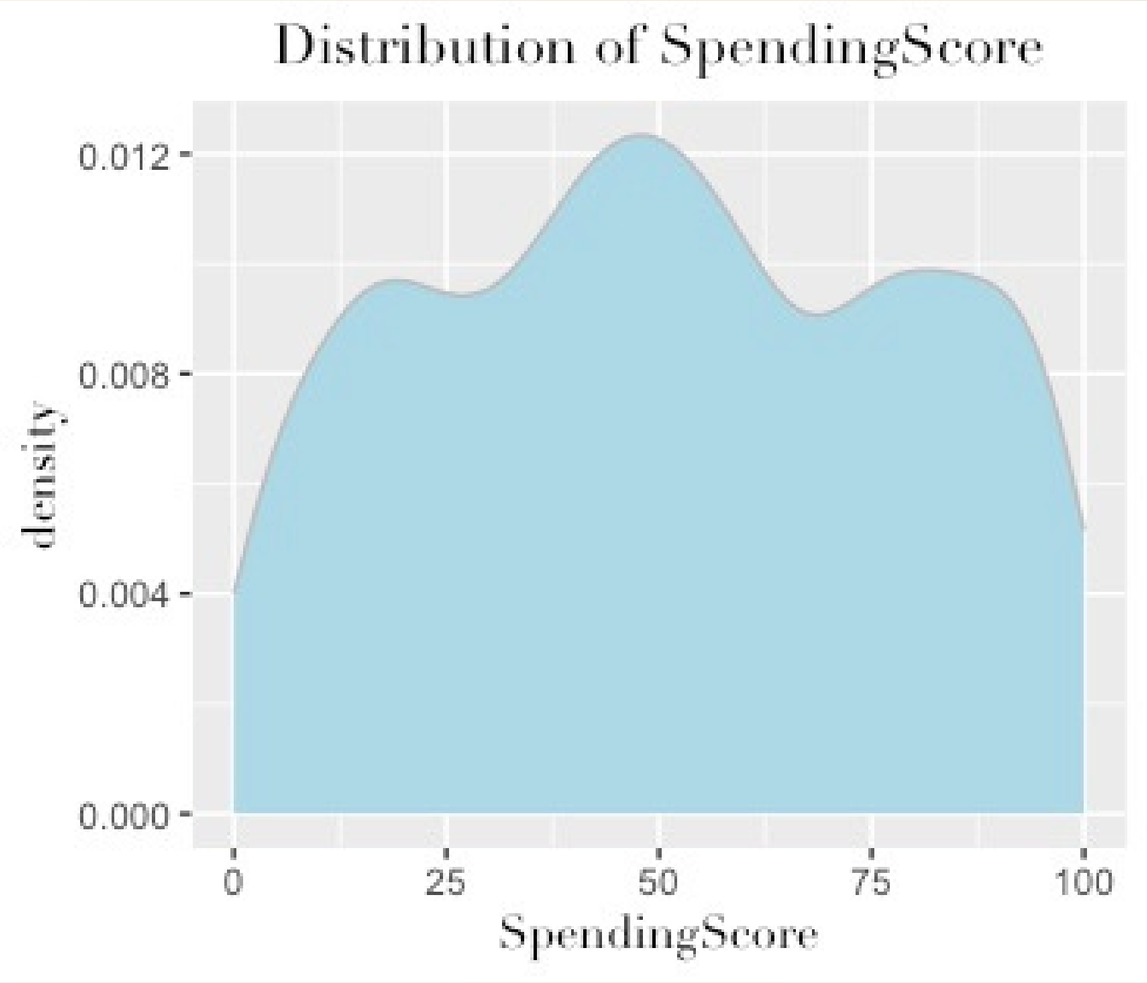
The median output of all consumers is 110,045 yuan; the first and fourth largest distributions are 74,572 yuan and 149,093 yuan respectively. This indicates that consumers' output is mainly distributed between 70,000 and 140,000 yuan.

Continuous Data: Age



The right figure shows that the age distribution in the sample is relatively consistent, and the median and mean are both 48 years old. It can be inferred that the age data may follow a symmetrical distribution, with the central tendency of age around 48 years old.

Continuous Data: Spending Score

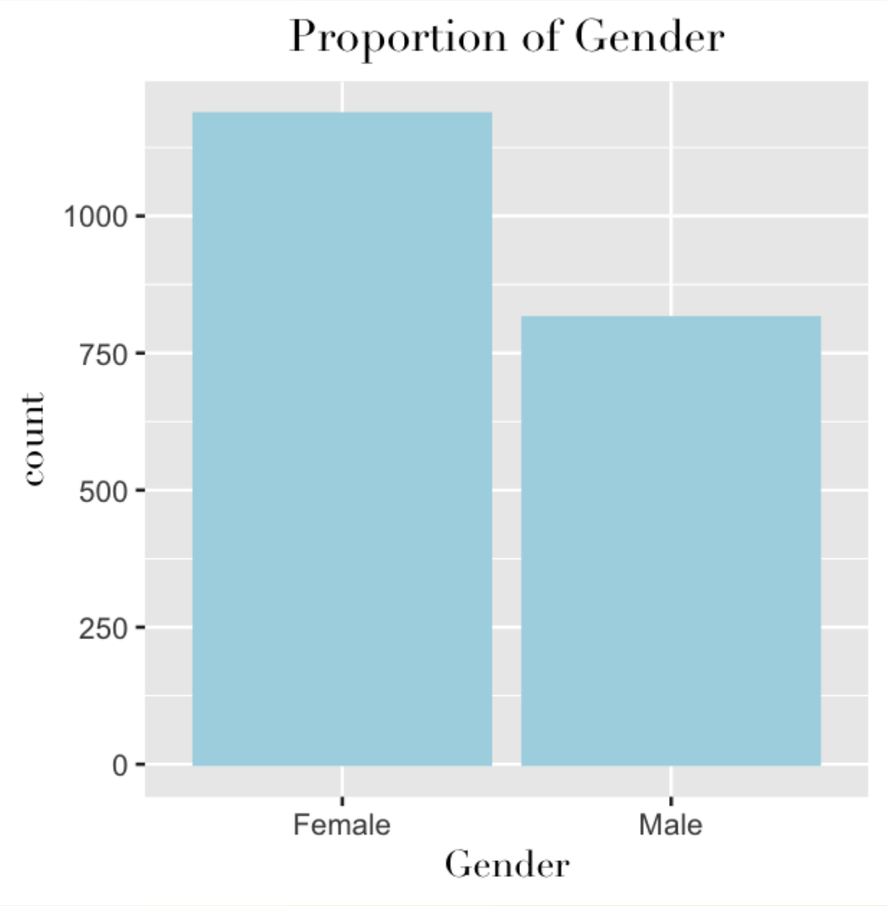


The consumer spending power score in the sample data ranged from 0 to 100. The median, mean, and mode were all 50, indicating that consumers' spending power was relatively concentrated and mainly distributed between 28 and 75.

03
Statistical Inference

- 性別、職業交叉消費能力
- 所得、年齡與消費能力的回歸分析
- 綜合變數分析

Gender-based Cross-tabulation of Spending Power (1/4)



Based on the descriptive statistics provided above, it can be observed that within this sample of 2,000 observations, the proportion of females is higher than that of males. We now intend to use this sample to make the following inferences regarding the population:

1. In the population, does the proportion of females exceed half?
2. In the population, is there a difference between the spending power of females and males?

Gender-based Cross-tabulation of Spending Power（2/4）

In the population, does the proportion of females exceed half?

Step1	H0: p<=0.5	H1: p>0.5
Step2	alpha=5%, conf.level=0.95	
Step3	One sample Proportion Test	
Step4	When p-value<0.05, reject H0	
Step5	p-value<2.2e-16	
Summary	由於p-value<alpha，故拒絕H0。女性比例顯著較男生高。	

Gender-based Cross-tabulation of Spending Power（3/4）

推論2：In the population, is there a difference between the spending power of females and males?

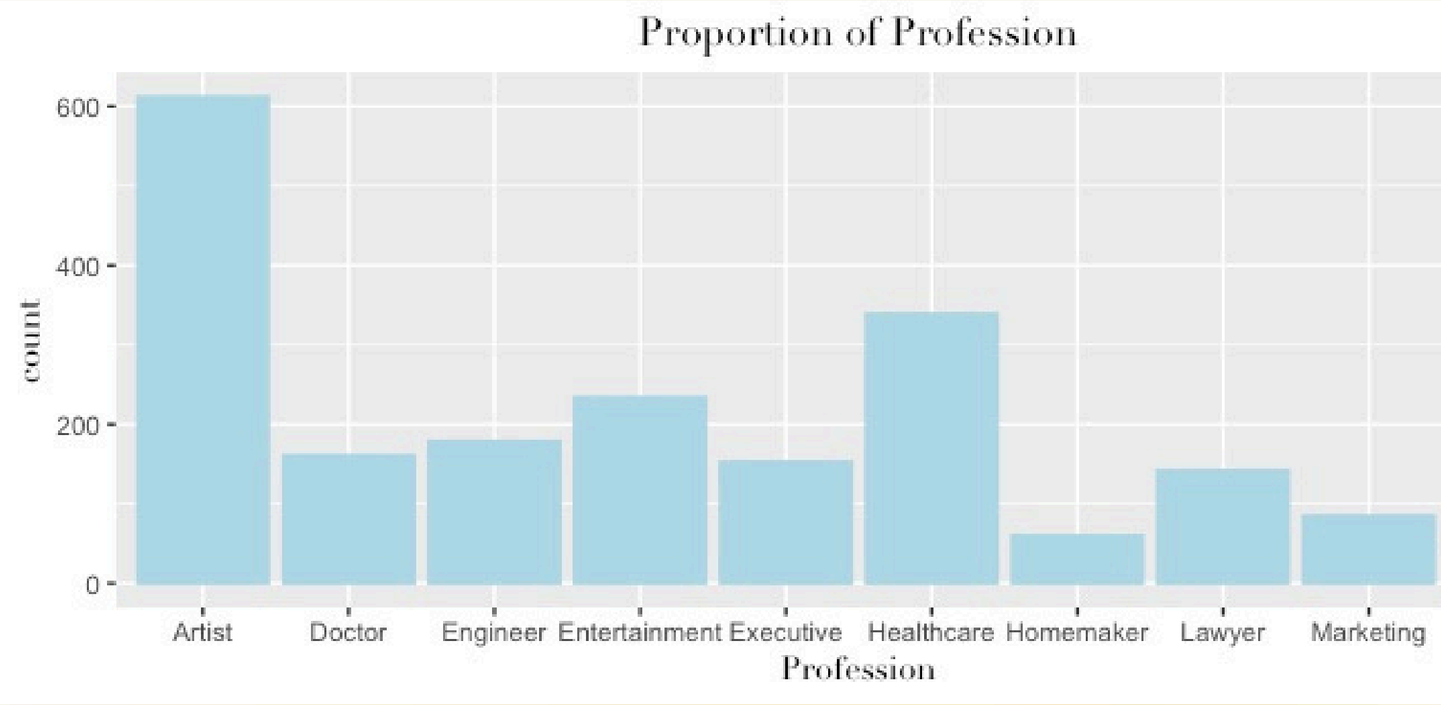
Step1	H0: σ男=σ女(ratio=1)	H1: σ男≠σ女(ratio≠1)
Step2	alpha=5%, conf.level=0.95	
Step3	F test	
Step4	When p-value<0.05, reject H0	
Step5	p-value<0.8065	
Summary	由於p-value>alpha，故不拒絕H0。得σ男=σ女	

Gender-based Cross-tabulation of Spending Power（4/4）

推論2：In the population, is there a difference between the spending power of females and males?

Step1	H0: μ男=μ女	H1: μ男≠μ女
Step2	alpha=5%, conf.level=0.95	
Step3	Two sample T test with equal Variance	
Step4	When p-value<0.05, reject H0	
Step5	p-value<0.9812	
Summary	由於p-value>alpha，故不拒絕H0。得μ男=μ女	

Cross-analysis of Spending Power by Occupation (1/3)



Descriptive statistics reveal that there are more than nine distinct occupations within this sample. We now intend to use this sample to make the following inferences:

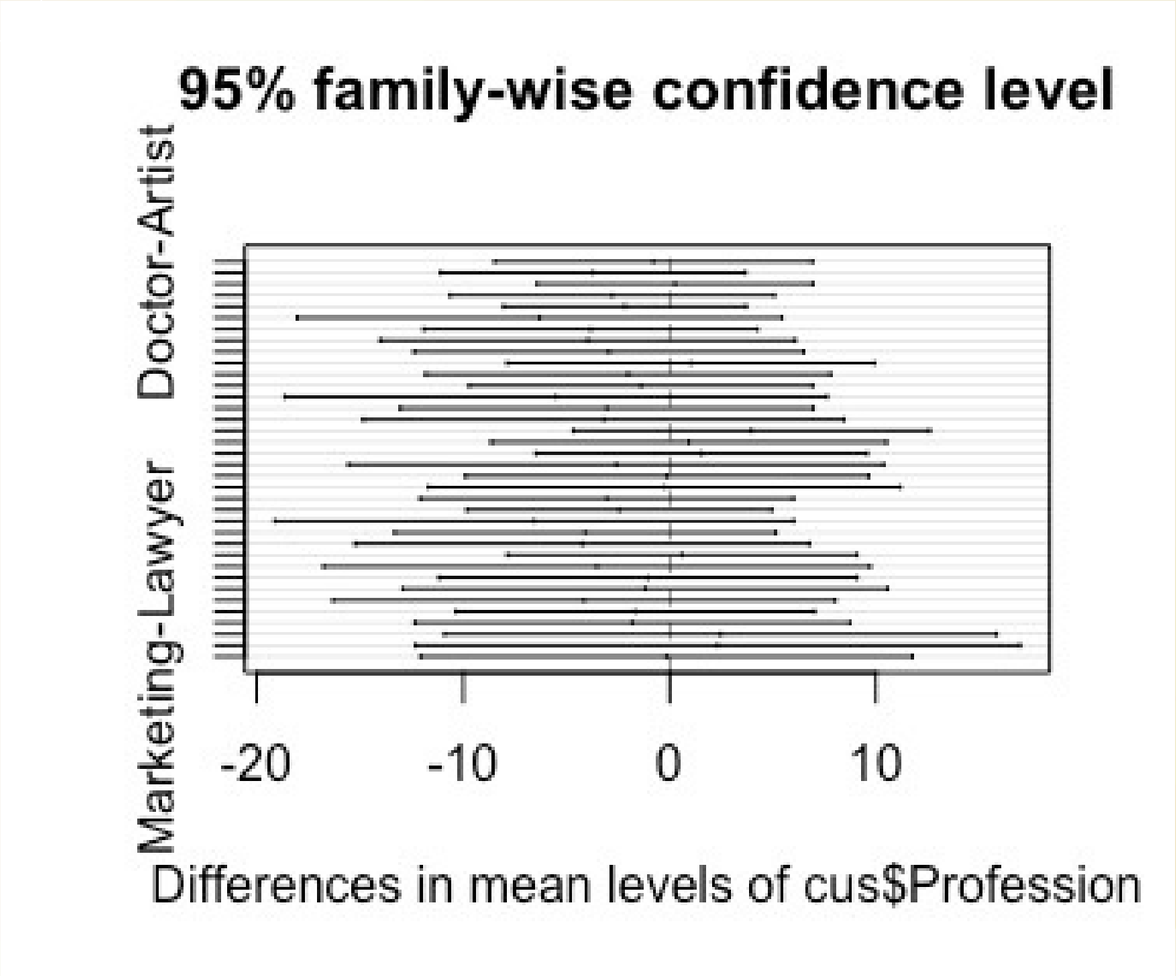
1. Do different occupations influence spending power?
2. If such an influence exists, which specific occupations exhibit significantly higher or lower spending power?

Cross-analysis of Spending Power by Occupation（2/3）

推論1：Do different occupations influence spending power?

Step1	H0: 所有職業消費能力均等	H1: H0為錯誤
Step2	alpha=5%, conf.level=0.95	
Step3	One way ANOVA	
Step4	When p-value<0.05, reject H0	
Step5	p-value<0.454	
Summary	由於p-value>alpha，故不拒絕H0。所有職業消費能力均等	

Cross-analysis of Spending Power by Occupation（3/3）

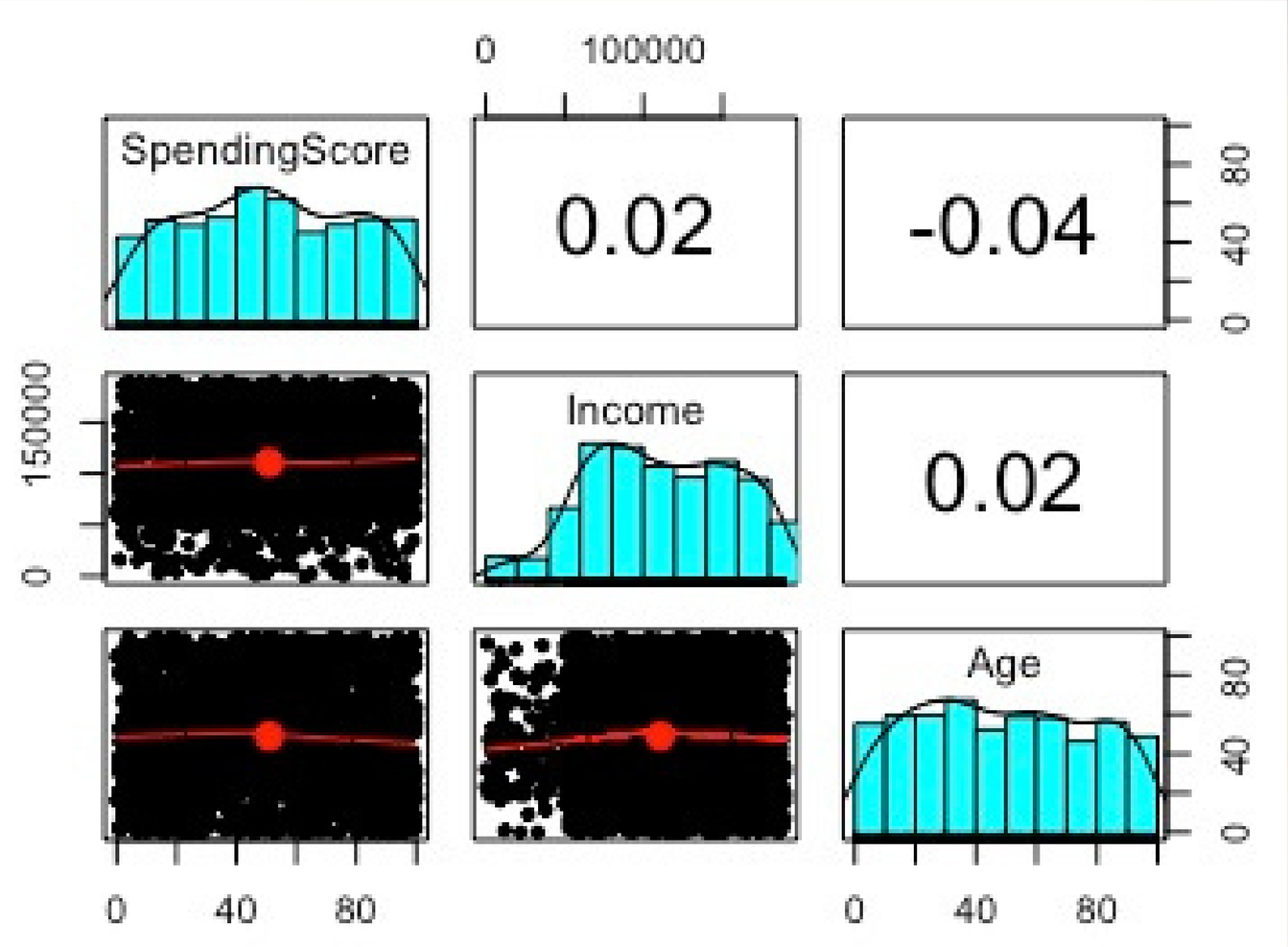


推論2：If such an influence exists (noting that the aforementioned ANOVA indicated no significant effect), which specific occupations exhibit significantly higher or lower spending power?

Post-hoc tests (Tukey’s HSD) can be employed to re-verify the validity of the conclusions derived from the aforementioned One-way ANOVA.

As illustrated in the figure on the left, the Tukey’s HSD confidence intervals for all pairwise occupational comparisons include zero. This indicates that there are no significant differences in spending power across various occupations, thereby confirming the accuracy of the previous One-way ANOVA results.

Income, Age and Spending Power



The scatter plot shows that:

- The correlation coefficient between age and spending power is 0.02, indicating a weak positive correlation.
- The correlation coefficient between age and spending power is -0.04, indicating a weak negative correlation.

Simple Regression of Income and Spending Power

49.39+1.423e-05xIncome	
β Pr(>|t|)	0.298
Adjusted R：	4.263e-05
F-statistic：P-value	0.2977

Simple Regression of Age and Spending Power

52.97330-0.0407xAge	
β Pr(>|t|)	0.0616
Adjusted R：	0.001247
F-statistic：P-value	0.06163

Simple Regression of Integrated Variables Discussion(1/2)

Income & Age

5.260e+01 + 1.729e-05xIncome - 3.834e-02xAge

Gender

Baseline = Female

GenderMale	0.2012

Simple Regression of Integrated Variables Discussion (2/2)

Profession, BaseLine = Artist

ProfessionDoctor	-0.9309
ProfessionEngineer	-3.529
ProfessionEntertainment	0.3018
ProfessionExecutive	-2.965
ProfessionHealthcare	-2.281
ProfessionHomemaker	-6.428
ProfessionLawyer	-3.908
ProfessionMarketing	-4.068

Prediction

- Profession : Doctor
- Genger : Male
- Age : 25
- Income : 50,000

51.77

48.75

- Profession : Lawyer
- Genger : Female
- Age : 30
- Income : 70,000

04 Conclusion

- 結論敘述

Conclusion

- Based on the previous ANOVA results, neither gender nor occupation shows a significant relationship with spending power. Furthermore, the simple linear regression analysis and its corresponding scatter plots indicate that age and income also lack a substantial correlation with spending capacity.
- In conclusion, none of the four variables analyzed in this dataset are significantly correlated with consumer spending power. This suggests that the current model should not be heavily relied upon for predicting consumer behavior. Future research should explore alternative variables to gain a deeper and more accurate understanding of the factors influencing spending power.