Assginment 1 Report, DATA 400

Intructor: Dr. Spectrum Han

Yohen Thounaojam, 56112204

Introduction

This report is a part of Assignment 1 of DATA 400, 2021W2 at UBCO. It is composed of two parts (1 report each). Please refer to the Table of Contents for more information.

Other Information

- All R code used in this report can be found in the .Rmd file in the submission. Moreover, for the relevant portions of the report, the code has been *'echoed'* back into this PDF.

Q1: Household Data Analysis

Page 41

Introduction

The data used in this analysis was collected from a survey of household expenditure. *Table 1.1* is 4 of 40 rows of expenditure of 20 single men and 20 single women in four commodity groups. The units of expenditure are in Hong Kong Dollars.

```
##    housing food goods service gender
## 19     382   77   230     147 female
## 20    1090   59   313     177 female
## 21     497  591   153     291   male
## 22     839  942   302     365   male
```

Table 1.1 4/40 rows of Household Expenditure data for single men and women in HK Dollars.

Objective

The aim of the survey was to investigate how the division of household expenditure between the four commodity groups depends on total expenditure and to find out whether this relationship differs between females and males.

Exploratory Data Analysis and Preparation

```
##     housing           food             goods           service
##  Min.   : 184.0   Min.   :  47.00   Min.   :   6.0   Min.   :  20.0
##  1st Qu.: 493.2   1st Qu.:  76.25   1st Qu.: 127.8   1st Qu.: 139.0
##  Median : 768.0   Median : 268.00   Median : 294.5   Median : 262.0
##  Mean   : 828.4   Mean   : 435.20   Mean   : 873.7   Mean   : 460.6
##  3rd Qu.:1033.5   3rd Qu.: 768.25   3rd Qu.: 948.2   3rd Qu.: 452.8
##  Max.   :1981.0   Max.   :1308.00   Max.   :6471.0   Max.   :2063.0
```

Table 1.2 Summary of data in Household dataset.

We also know that we are concerned with the total expenses too. So, we will add a new column of total expenses.

```
data("household", package = "HSAUR3")
household$total <- rowSums(household[0:4])
household[19:22,0:6]
```

```
##    housing food goods service gender total
## 19     382   77   230     147 female   836
## 20    1090   59   313     177 female  1639
## 21     497  591   153     291   male  1532
## 22     839  942   302     365   male  2448
```

Table 1.3 4/40 rows of Household Expenditure data after adding Total Expenditure Column (in HK$).

Method

The key ideas behind this survey were: 1. Relationship of each category to the total values. 2. Similarities/Differences of the above trend between males and females.

Plot 1

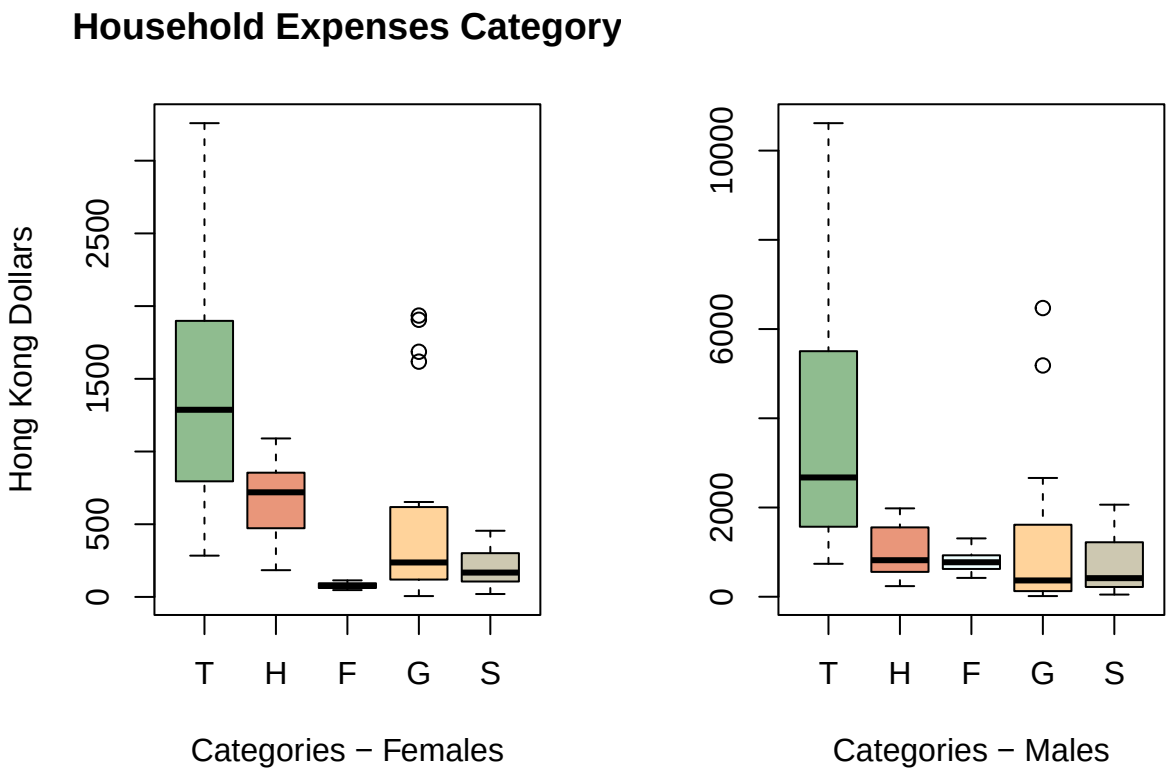
As seen below, there are two plots. Both plots have: Categories on the x-axis and Expense in HK$ on the y-axis.

Each plot represents all the female and male data respectively. This method is useful because in this way, we are able to compare the categories to the total expenses and also compare the two surveyed genders side by side.

Plot 2

In Plot 2, a scatterplot matrix is presented for all the categories to be compared to each other. More explanation later.

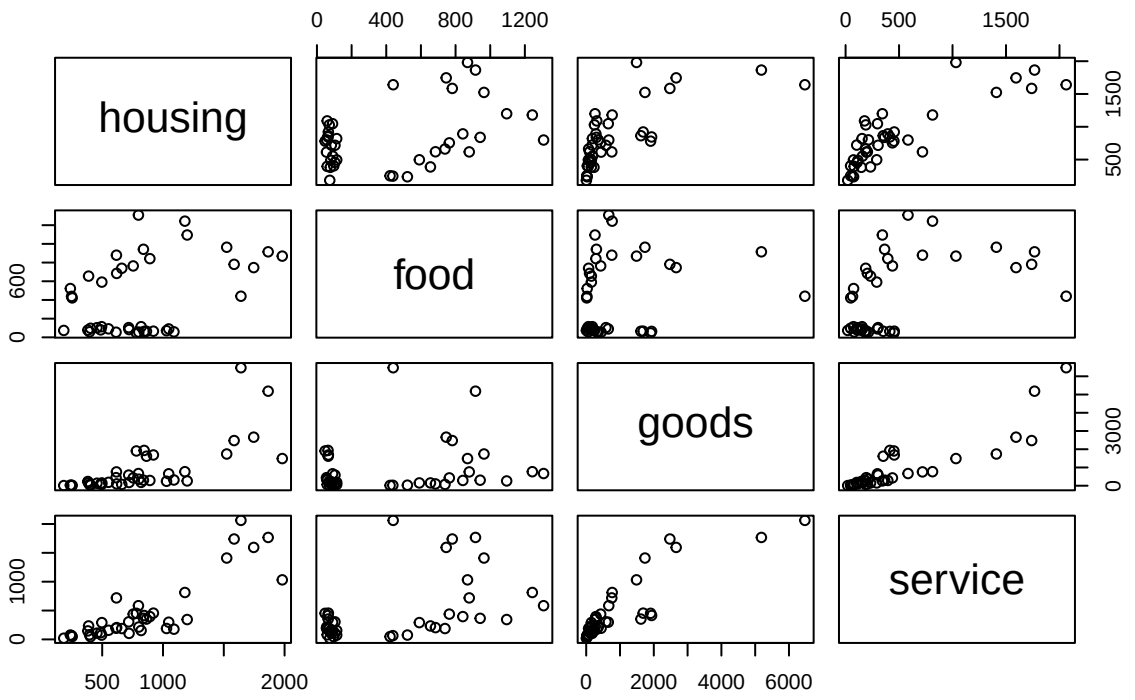
Results



Abbreviation	T	H	F	G	S
Category	Total	Housing	Food	Goods	Service

Plot 1: (A) Categories - Females, (B) Categories - Males

For aesthetic purposes, the categories were abbreviated.



Plot 2: Scatterplot displaying relationships between different expense categories (Both genders combined).

Relationship between categories over both genders

Plot 2 presents us a good visualization of the spending nature/habits of single men and women combined. Yes, we are interested in the differences between the two genders and we will look at that in the next sections of this report.

Here, we can see that most relationships are positive with a linear nature; meaning, for example, the more they spend on housing, the more likely they are to spend more on goods.

One interesting thing we notice is that in the Housing~Food plot with Housing in the x-axis, after a certain increase in housing expense, the food prices either drop or stay consistent. This may be due to better kitchen facilities in the costlier homes, making the single men and women cook more at home, which is cheaper than going out.

Another interesting plot is on the top right. As Housing prices increase, we see a linear increase in service expense. This is clear that with costlier housing come costlier services. This is helpful to explain to potential homeowners/renters that they have not have planned for unforeseen service fees.

Expenses for Females

In Plot 1 (A), we see the box plot of the distribution of the household expenses for females in HK$. Here, the highest expense compared to the total is that of *Goods*, followed by *Housing* and *Service* with minimal spending in *Food*.

Approximate Median Expenses (In Order)

1. Total : HK$ 1300
2. Housing : HK$ 750, Goods : HK$ 300, Service : HK$ 250, HK$ 100

Expenses for Males

In Plot 1 (B), we see the box plot of the distribution of the household expenses for males in HK$. Here, the highest expense compared to the total is that of *Food* and *Housing* followed by *Service* and then *Goods*.

Approximate Median Expenses (In Order)

1. Total : HK$ 2800
2. Housing : HK$ 1000, Food : HK$ 1000, Service : HK$ 600, Goods : HK$ 500

Comparison: Female and Male Expenses

We see that the median total expense for males is significantly higher than that of females. A big reason for that is the high spending in food in males' expenses. While females have a median expense of approximately HK$ 100, males have a surging HK$ 1000 food expense; which is ten times.

That being said, food is not the only contributing factor to the high total expense of males in the survey. We notice a higher expense in all other categories as well.

Discussions and Conclusion

First, we have established that high expenses for males is mostly caused by high spending in food. Next, when we look at the housing expense, we also notice that it is higher for males.

Here we have a possible hypothesis that since males are spending more on housing, their residences must be located in convenient and happening places in Hong Kong (e.g. near the Central District). This presents a much higher chance of the males going out to nearby restaurants to eat which we can all agree usually costs more than cooking at home.

Similarly, the more expensive the housing is, we can naturally expect a higher cost in services as well.

Recommendations

1. The survey only covered 20 residents from each gender; this may introduce bias depending on which sampling method was employed. Alternatively, for a city as highly populated as Hong Kong, significantly more residents need to be surveyed to come to a solid conclusion.
2. Since where in the city the residents live and work may highly influence their expenses, the survey must include that data.

Q3: Suicides Data Analysis

Page: 44

Introduction

The data represented in this section of the report represents mortality rates per 100,00 from male suicides for a number of age groups and a number of countries.

The following table is the first 6 rows of the data. Please note that there are more coutries listed in the dataset.

```
##         A25.34 A35.44 A45.54 A55.64 A65.74
## Canada      22     27     31     34     24
## Israel       9     19     10     14     27
## Japan       22     19     21     31     49
## Austria     29     40     52     53     69
## France      16     25     36     47     56
## Germany     28     35     41     49     52
```

Table 3.1 6/15 rows of Suicides Data for Countries divided in 5 age groups.

Objective

The goal of this analysis is to extract any specific correlations between the age groups and the mortality rate for the given countries.

The question specifies that a box plot must be created for the analyses.

Exploratory Data Analysis and Preparation

```
##      A25.34          A35.44          A45.54         A55.64         A65.74
##  Min.   : 4.00   Min.   : 7.00   Min.   :10.0   Min.   :14.0   Min.   : 22.00
##  1st Qu.: 9.50   1st Qu.:16.00   1st Qu.:16.5   1st Qu.:19.0   1st Qu.: 27.00
##  Median :22.00   Median :25.00   Median :31.0   Median :33.0   Median : 35.00
##  Mean   :19.93   Mean   :26.33   Mean   :32.0   Mean   :36.4   Mean   : 42.27
##  3rd Qu.:27.00   3rd Qu.:34.50   3rd Qu.:41.0   3rd Qu.:49.5   3rd Qu.: 51.50
##  Max.   :48.00   Max.   :65.00   Max.   :84.0   Max.   :81.0   Max.   :107.00
```

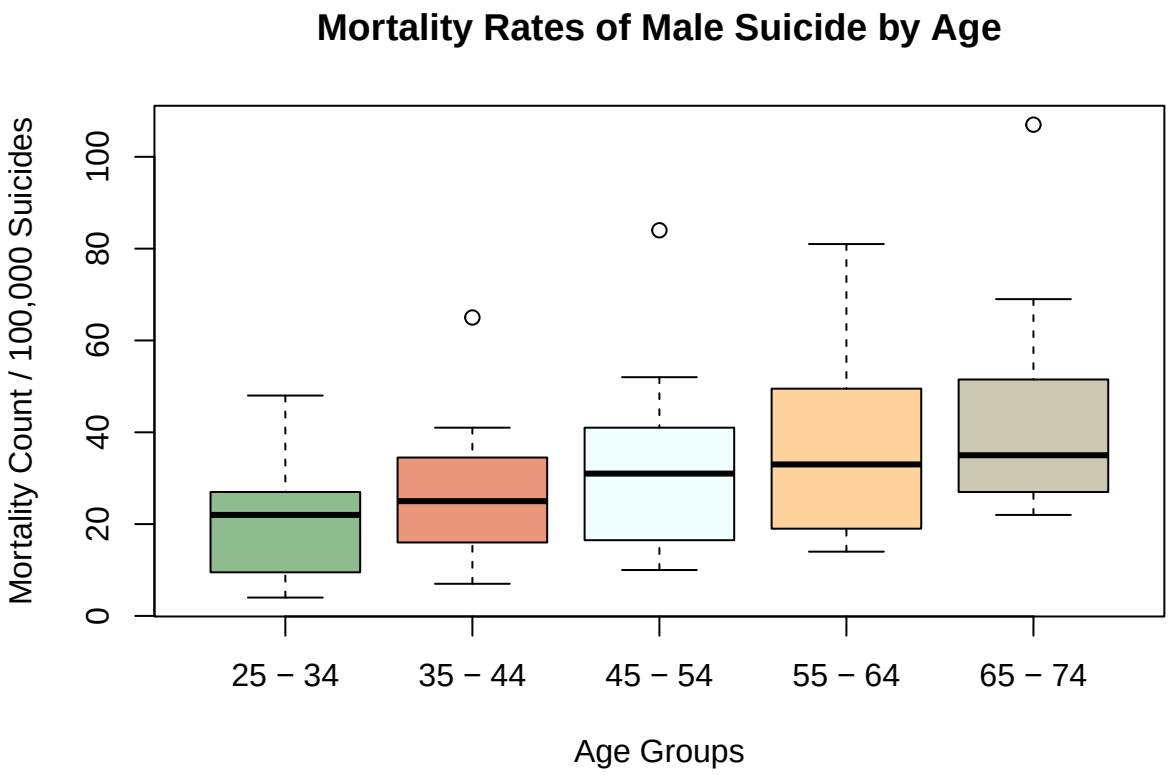
Table 3.2 Summary of data in Suicides2 dataset.

Methods

We will be using box plots, arranged side-by-side, to analyze and discuss what we see. The box plots will be constructed with the mortality rate on the x-axis and the age groups on the y-axis. Each mortality rate data point represented by the box plot is associated to a country.

Please see box plot below.

Results



**Plot 3: Mortality rates of map suisides2. “ ”

Discussions and Conclusion

The following is a detailed yet short analysis of the plot above.

Age Group	Median	Max.	Min.	Outliers
25-34	21	48	04	None
35-44	23	40	06	01
45-54	30	51	10	01
55-64	32	80	15	None
65-74	39	70	21	01

Here, the first clear trend is that the more their age is, the more likely are to count towards the mortality rate. I suspect that this is because the older men are more sure of what they want while the men after 34 are unsure. This confusion may subconsciously cause a failure.

Acknowledgements

Data Source: HSAUR3 | *Link: Documentation*

References

Question from the A Handbook of Statistical Analyses Using R | *Link*