Household and Suicides Analysis Report

Ryota Nishida

Introduction

This report is written with the objective of describing and exploring data frames found within the book "A Handbook of Statistical Analyses Using R" by Brian S. Everitt and Torsten Hothorn. The data is retrieved from the HSAUR3 package found within R Studio. The data will be explored numerically and graphically with a focus on two data frames, one of which will include a numerical response variable and another which will include a categorical response variable.

1.1 Data Description

The data sets that will be analyzed from the HSAUR3 package are the **household** and **suicides2** data sets.

The household data set showcases data collected from a survey of household expenditure and gives the expenditure of 20 single men and 20 single women on four commodity groups. The units of expenditure are Hong Kong dollars, and the four commodity groups are housing(housing, including fuel and light),food(foodstuffs, including alcohol and tobacco), goods(other goods, including clothing, footwear and durable goods), and services(services, including transport and vehicles).For the household data set, the aim will be 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 for men and women (Hothorn & Everitt, 2014 [1]).

The suicides2 data set showcases the mortality rates per 100,000 from male suicides for a number of age groups and a number of countries. The aim for the suicides2 data set will be to numerically and graphically examine the relationship between mortality and different age groups (Hothorn & Everitt, 2014 [1]).

Table 1: Household Data

housing	food	goods	service	gender
820	114	183	154	female
184	74	6	20	female
921	66	1686	455	female
488	80	103	115	female
721	83	176	104	female
614	55	441	193	female
801	56	357	214	female
396	59	61	80	female
864	65	1618	352	female
845	64	1935	414	female
404	97	33	47	female
781	47	1906	452	female
457	103	136	108	female
1029	71	244	189	female
1047	90	653	298	female
552	91	185	158	female
718	104	583	304	female
495	114	65	74	female
382	77	230	147	female
1090	59	313	177	female
497	591	153	291	male
839	942	302	365	male
798	1308	668	584	male
892	842	287	395	male
1585	781	2476	1740	male
755	764	428	438	male
388	655	153	233	male
617	879	757	719	male
248	438	22	65	male
1641	440	6471	2063	male
1180	1243	768	813	male
619	684	99	204	male
253	422	15	48	male
661	739	71	188	male
1981	869	1489	1032	male
1746	746	2662	1594	male
1865	915	5184	1767	male
238	522	29	75	male
1199	1095	261	344	male
1524	964	1739	1410	male

Table 2: Suicides Data

	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
Hungary	48	65	84	81	107
Italy	7	8	11	18	27
Netherlands	8	11	18	20	28
Poland	26	29	36	32	28
Spain	4	7	10	16	22
Sweden	28	41	46	51	35
Switzerland	22	34	41	50	51
UK	10	13	15	17	22
USA	20	22	28	33	37

Exploratory Data Analysis

Household Data Analysis

Household data summary

```
##     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  
##     gender  
##  female:20  
##  male  :20  
##             
##             
##             
##             
```

1. *Above is a numerical summary of the household data frame*

Housing linear relationship summary

```
##
## Call:
## lm(formula = housing ~ food + goods + service, data = household)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -420.97 -173.03  -17.64  125.41  707.64
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 492.07243   62.17147   7.915 2.15e-09 ***
## food         -0.04457    0.14382  -0.310 0.758434
## goods        -0.04910    0.07673  -0.640 0.526285
## service       0.86543    0.22295   3.882 0.000425 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 253.1 on 36 degrees of freedom
## Multiple R-squared:  0.7233, Adjusted R-squared:  0.7002
## F-statistic: 31.36 on 3 and 36 DF,  p-value: 3.803e-10
```

2. Above is a numerical summary of the linear relationship between housing and different expenditure types.

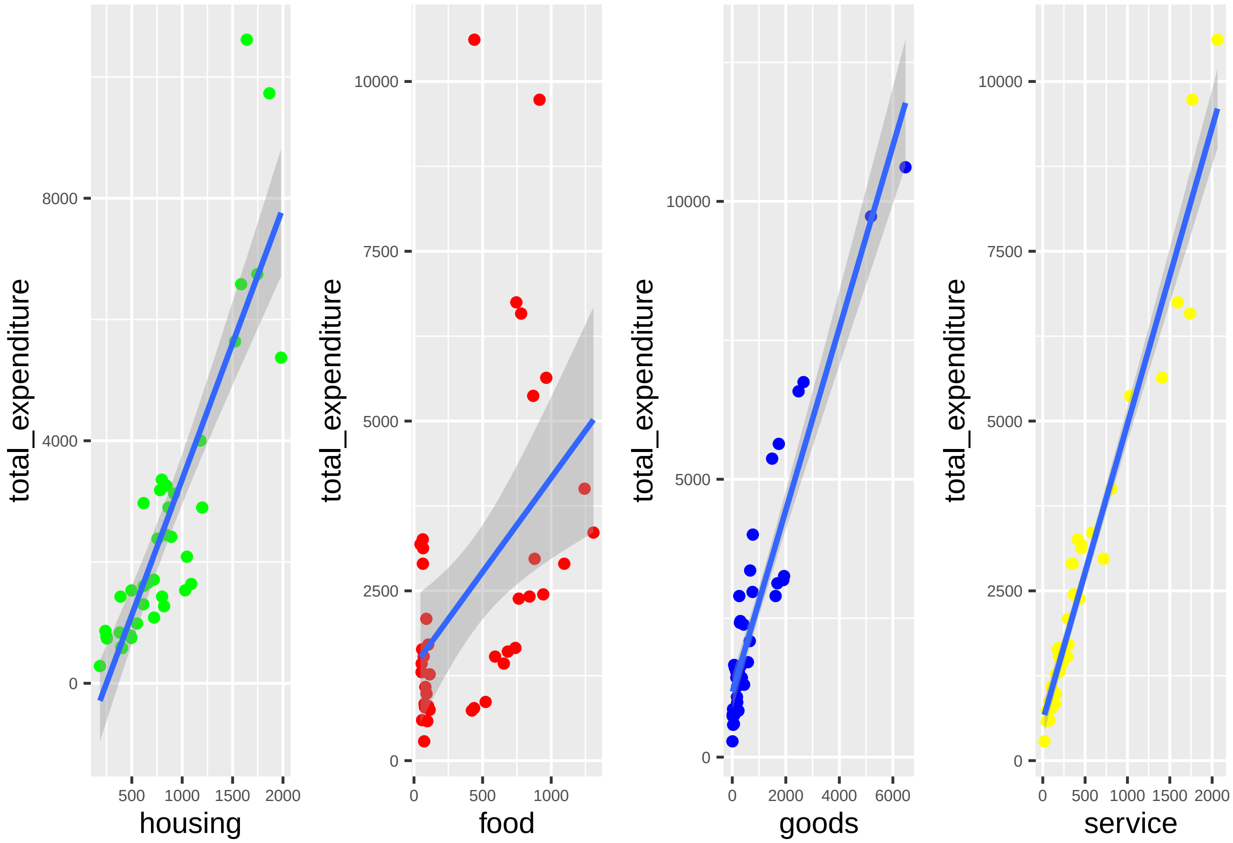

Figure 1: Scatter plots showcasing the relationship between total expenditure and different expenditure types

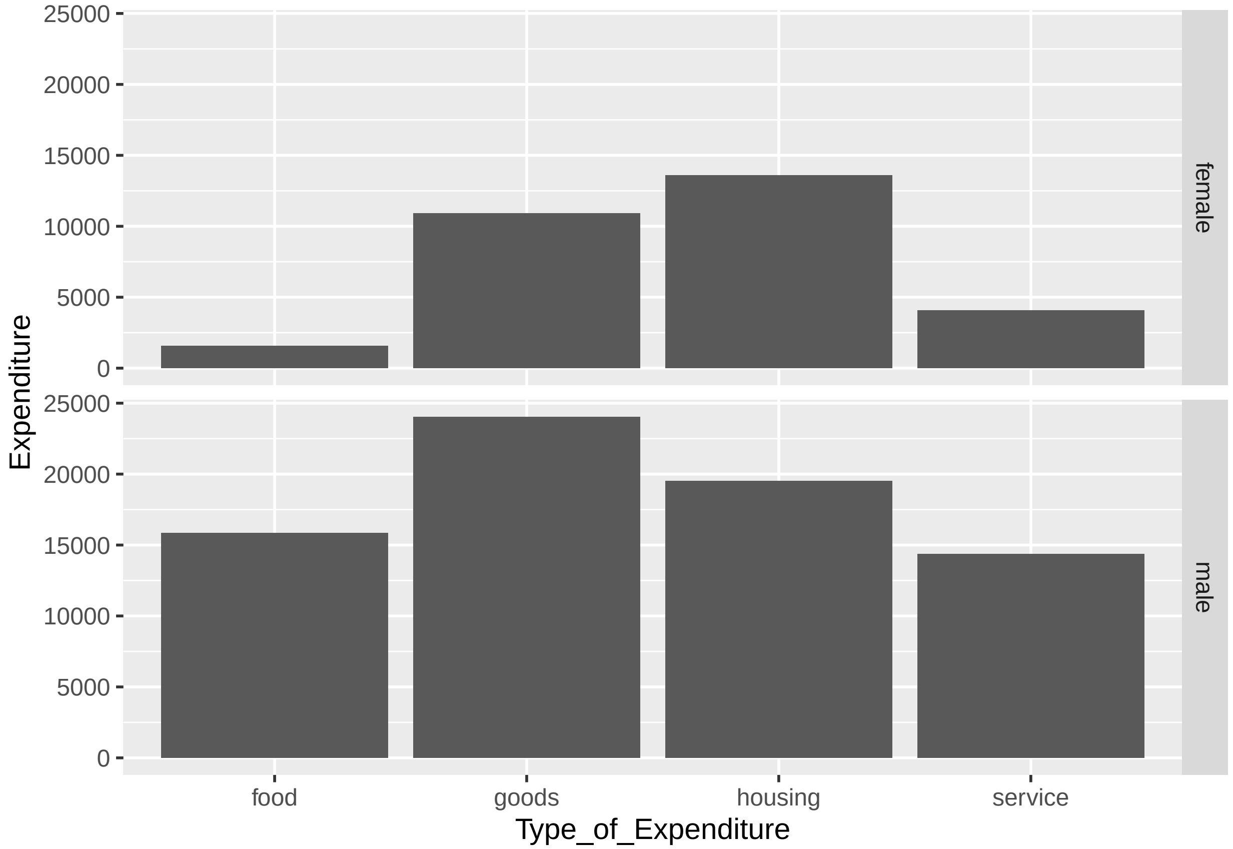

Figure 2: Faceted bar chart showing the difference between how males and females distribute their expenditures across different types of expenditure

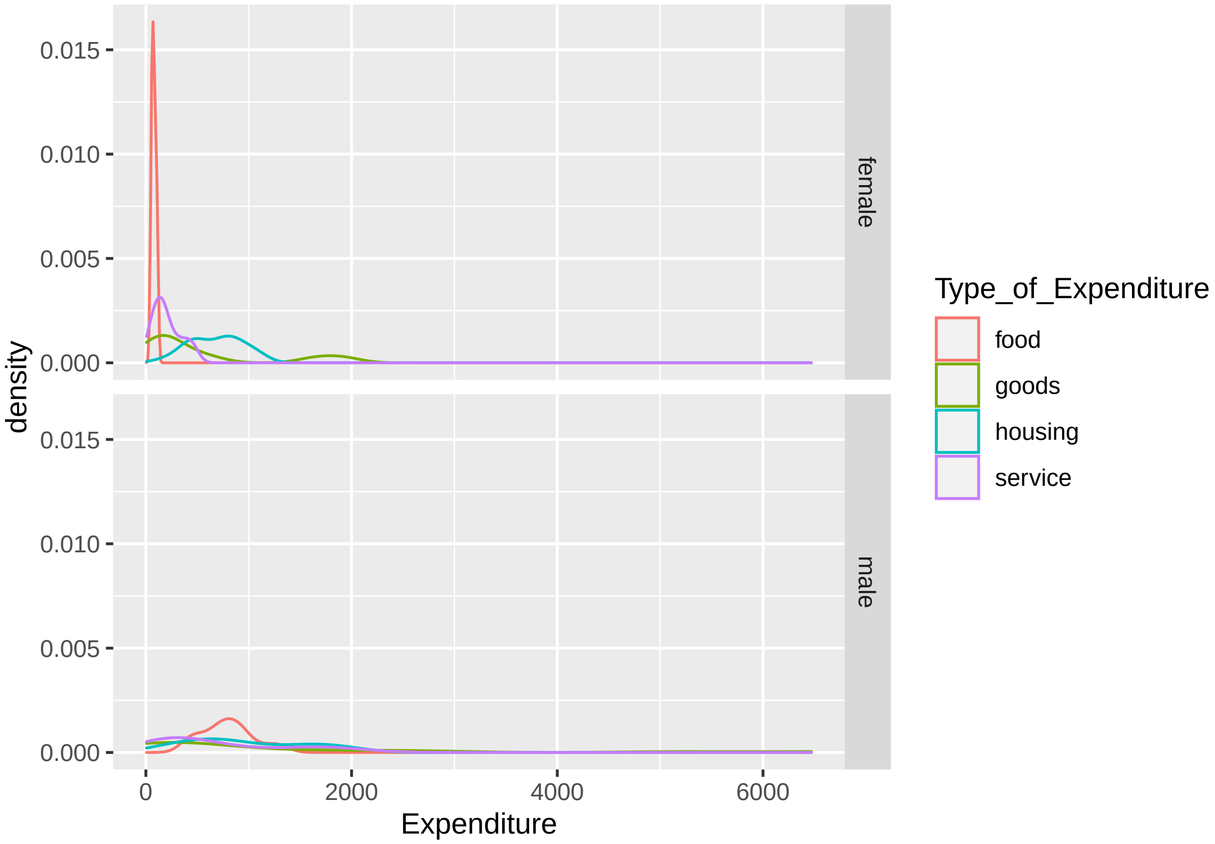

Figure 3: Density curves of each expenditure by gender

Suicides2 Data Analysis

Suicides2 numerical summary

```
##      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
```

3. Above is a numerical summary of the suicides2 data frame.

Suicides2 linear regression summary

```
##
## Call:
## lm(formula = Suicides ~ Countries + Age_Group, data = suicides3)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -17.5467  -2.4133  -0.5467   4.0533  19.1200
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)            37.147      3.496  10.627 4.82e-15 ***
## CountriesCanada       -21.000      4.392  -4.781 1.30e-05 ***
## CountriesFrance       -12.600      4.392  -2.869   0.0058 **
## CountriesGermany       -7.600      4.392  -1.730   0.0891 .
## CountriesHungary       28.400      4.392   6.466 2.63e-08 ***
## CountriesIsrael       -32.800      4.392  -7.467 5.84e-10 ***
## CountriesItaly        -34.400      4.392  -7.832 1.46e-10 ***
## CountriesJapan        -20.200      4.392  -4.599 2.47e-05 ***
## CountriesNetherlands  -31.600      4.392  -7.194 1.65e-09 ***
## CountriesPoland       -18.400      4.392  -4.189   0.0001 ***
## CountriesSpain        -36.800      4.392  -8.378 1.85e-11 ***
## CountriesSweden        -8.400      4.392  -1.912   0.0609 .
## CountriesSwitzerland   -9.000      4.392  -2.049   0.0452 *
## CountriesUK           -33.200      4.392  -7.559 4.13e-10 ***
## CountriesUSA          -20.600      4.392  -4.690 1.80e-05 ***
## Age_GroupA35.44         6.400      2.536   2.524   0.0145 *
## Age_GroupA45.54        12.067      2.536   4.758 1.41e-05 ***
## Age_GroupA55.64        16.467      2.536   6.493 2.37e-08 ***
## Age_GroupA65.74        22.333      2.536   8.807 3.71e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.945 on 56 degrees of freedom
## Multiple R-squared:  0.9018, Adjusted R-squared:  0.8702
## F-statistic: 28.55 on 18 and 56 DF,  p-value: < 2.2e-16
```

4. Above is a linear regression showcasing a summary of the relationship between suicides and independent variables such as countries and age groups.

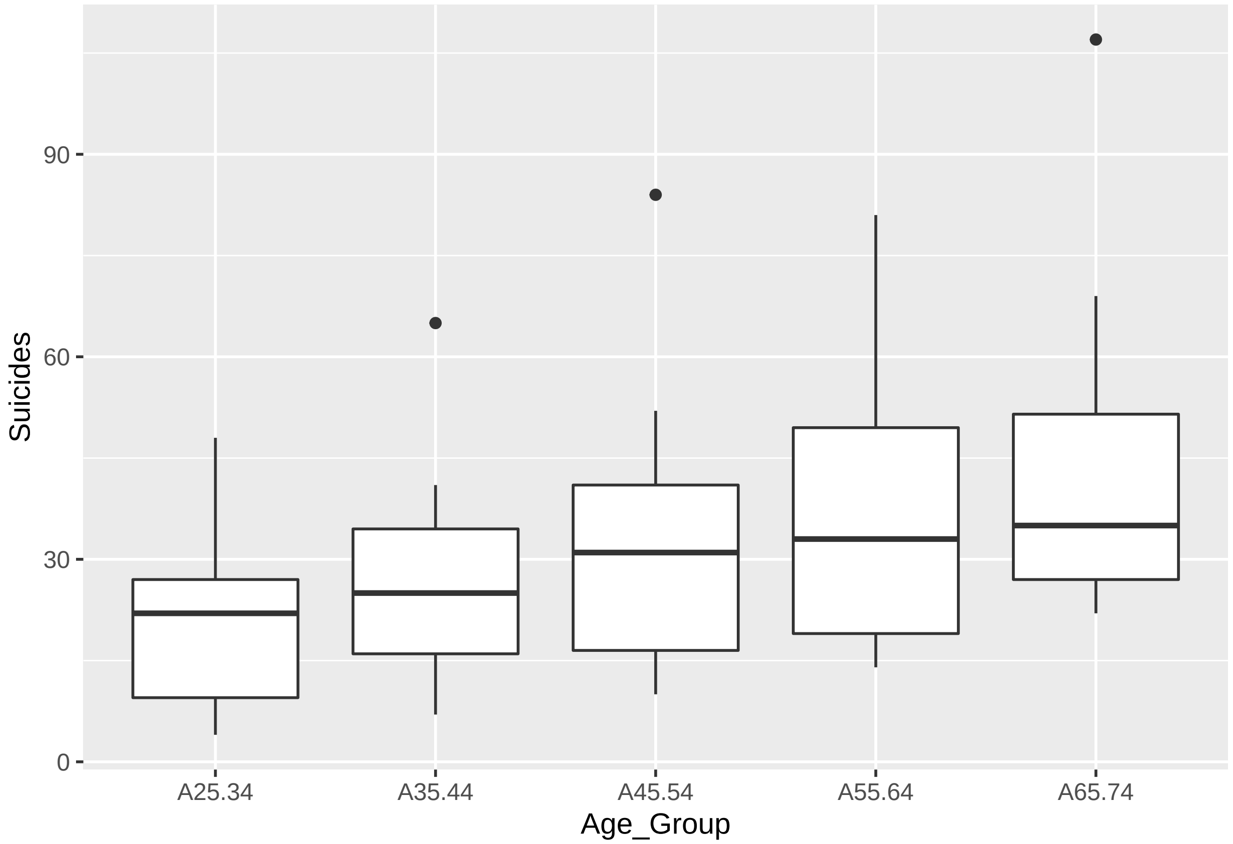

Figure 4: Sideby-side box plots for the data on suicides from different age groups of males

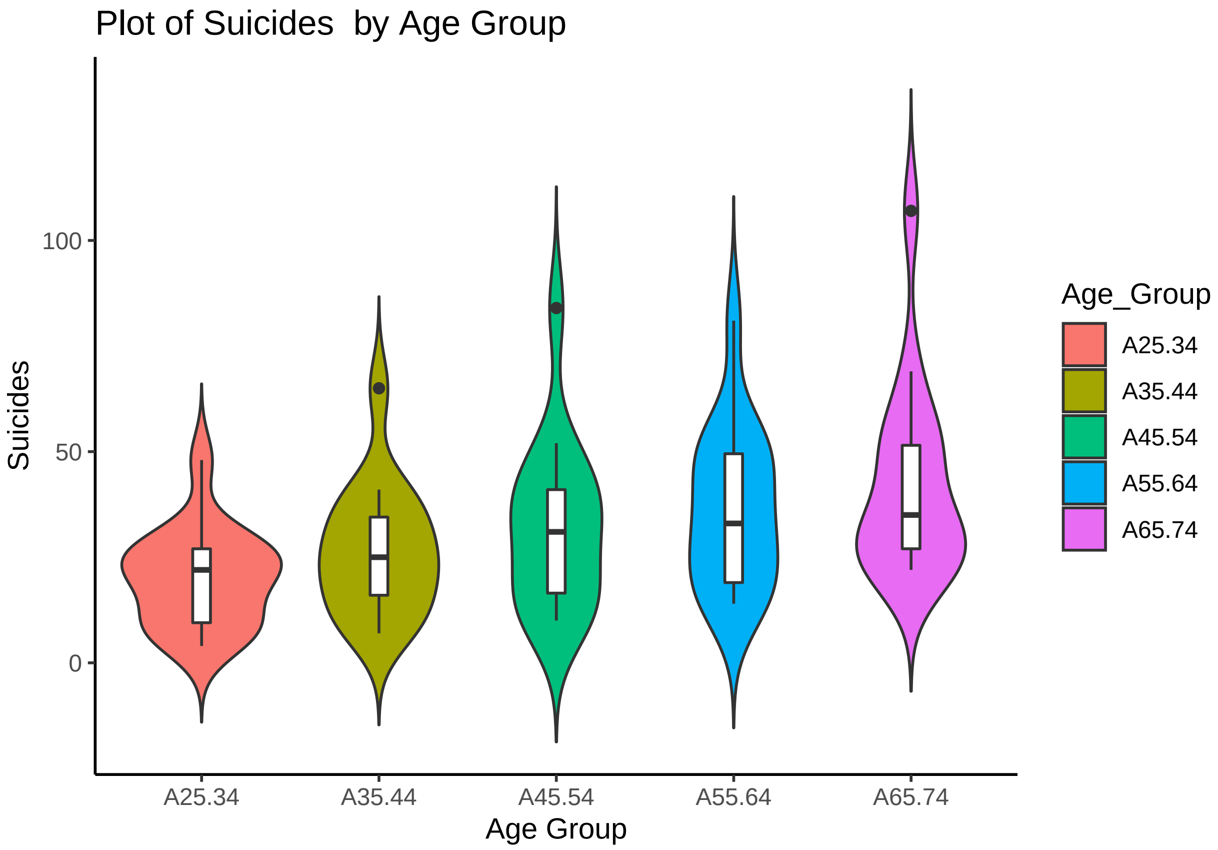

Figure 5: Violin plot for suicides by age groups

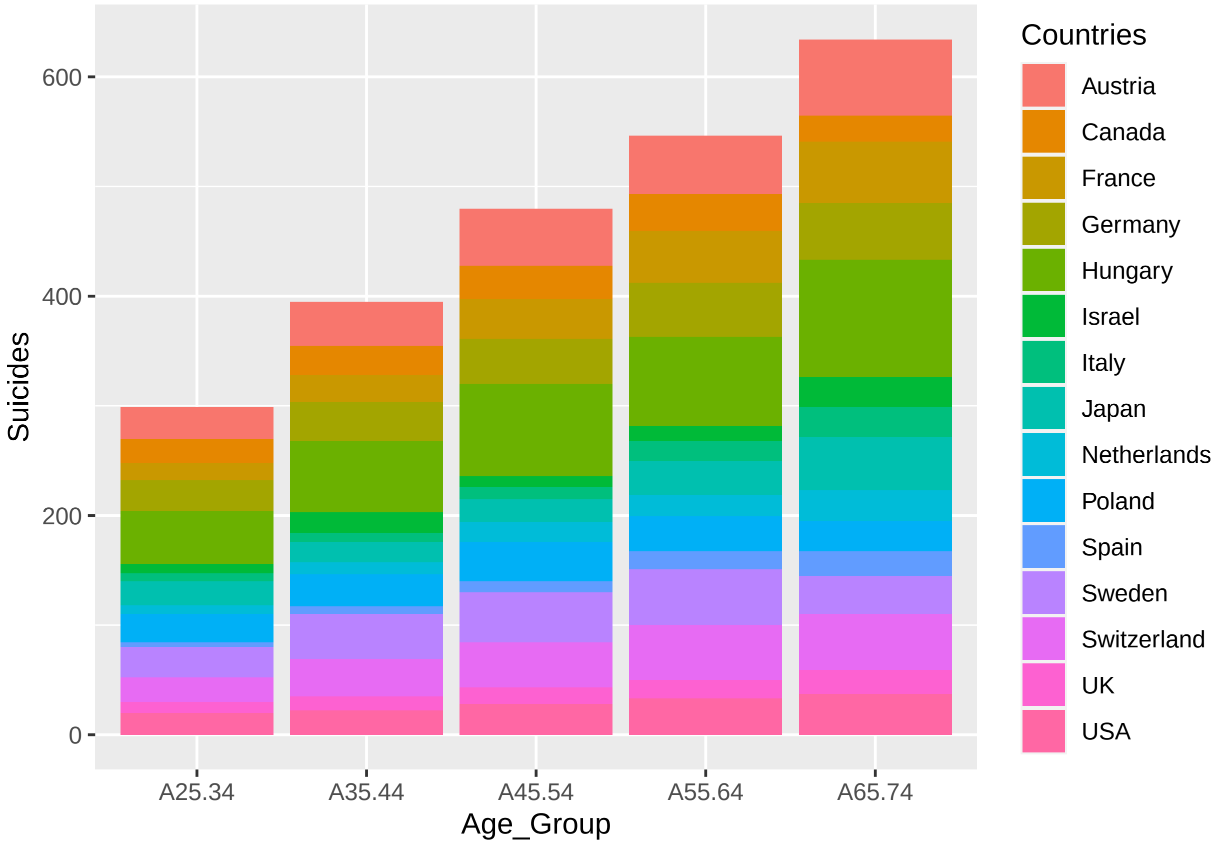

Figure 6: Bar chart showing suicide rates for different age groups in different countries

Methods

All statistical analysis, such as the use of linear regression, standard deviation, and t-value was done using the R Studio program. The **household** and **suicides2** data frames were used as the data frames for analysis. An assortment of R Studio packages such as ggplot and Tidyr were used in order to reproduce graphs and numerical summaries such as scatter plots and bar charts.

Results of Analysis

Household data analysis

The numerical and graphical data exploration done in the previous section has brought forth different observations.

For instance, the household data summary shows that the minimum spending for the expenditure types are 184(housing),47(food),6(goods), and 20(service). The maximum spending for the expenditure types are 1981(housing),1308(food),6471(goods), and 2063(service). Goods has the highest range of values and food has the lowest. There are 20 females and 20 males included in the data. The median values of the expenditure types are 768(housing), 268(food),294.5(goods), and 262(service). Most of the medians are within the 200-300 range, with the exception of the housing median value which is much higher. The mean spending of the expenditure types are 828.4(housing), 435.2(food),873.7(goods), and 460.6(service). The means for housing and goods are both between 800-900 and the means for food and service are both between 400-500.

The housing linear relationship summary showcases that there is a negative relationship between housing and food(ceteris paribus), and with housing and goods(ceteris paribus). However, there is a positive relationship between housing and service. Using a significance level of 0.05, only the results for service are statistically significant. The adjusted R SQUARED value of 0.7 shows that 70 percent of the variation in housing can be explained by the different expenditure types.

The scatter plots in figure 1 show that there is a positive relationship between service expenditure and total expenditure. When there is an increase in total expenditure, there is also an increase in spending with regards to service. Spending on housing and goods also increase with an increase in total expenditure, however they do not possess the same level of correlation as expenditure on service. Furthermore, the relationship between total expenditure and food expenditures cannot be determined because the data does not fit a line.

The bar graphs in figure 2 show that women spend the most money on housing and the least money on food. Men on the other hand, spend the most money on goods and the least money on services. Overall expenditure is higher for men than women.

The density curves in figure 3 also show that females spend less than males when it comes to expenditure on food. Additionally, there is more of an even distribution of housing expenditures for females than there is for males. The graph also shows that the expenditure for women and men is right-skewed.

Suicides2 data analysis

The numerical summary for suicides2 shows that the minimum suicides for the age groups are 4(A25.34),7(A35.44),10(45.54), 14(A55.64), and 22(A65.74). The maximum suicides for the age groups are 48(A25.34),65(A35.44),84(45.54), 81(A55.64), and 107(A65.74). The A65.74 age group has the highest range of values and the A25.34 age group has the lowest.The median suicide values of the age groups are 22(A25.34),25(A35.44),31(45.54), 33(A55.64),and 357(A65.74). Most of the medians are within the 30-35 range, with the exception of the A25.34 and A35.44 age groups which are in the 20-25 range. The mean values of suicides for the different age groups are 19.93(A25.34),26.33(A35.44),32(45.54), 36.4(A55.64),and 42.27(A65.74). The A45.54 and A55.64 age groups have similar means within the 30-40 range.

The suicides linear relationship summary showcases that there is a negative relationship between suicides and the regressors for different countries(ceteris paribus), with the exception of Hungary. However,there is a positive relationship between suicides and the age groups. Using a significance level of 0.05, all the regressors have a statistically significant relationship with suicides. The adjusted R SSQUARED value of 0.87 shows that 87 percent of the variation in suicides can be explained by the different country country and age group regressors.

The box plots in figure 4 show that the data for the A35.44,A45.54, and A65.74 age groups contain the presence of outliers. The A65.74 age group has the highest level of suicides, while the A25.34 age group has the least level of suicides. The A55.65 age group has the largest standard deviation while the A25.34 and A35.44 age groups have the lowest standard deviations.

The violin plots in figure 5 show that as the age groups increase, the number of suicides increase along with them. The tails of the violin plots show that the data for the age groups is positively skewed. The A65.74 age group has the longest tail distribution and the A25.34 age group has the shortest. The median for the A25.34 age group is pulled closer to the Q3, and that of the A65.74 age group is pulled closer to Q1; the rest are centered. The A45.54 and A55.64 age groups have elongated distributions without distinct peaks. The A25.34 and A65.74 age groups showcase a bimodal distribution and the A35.44 age group features a normal distribution.

The bar graphs in figure 6 also show that suicides increase as the age group increases. Furthermore, this is the case for most countries. The graph also shows that Hungary has the highest number of suicides for each age group compared to other countries and Spain has the lowest.

Discussion

Household data discussion

The household data analysis has shown that there is a clear relationship between service expenditure and total expenditure as can be seen in figure 1 . It has also shown that overall, males spend more than females across all expenditure types as is evident from figure 2. However, this may be because males have more to spend than females. This suggests that males may make more money than females and therefore have higher levels of disposable income available to them. The fact that females spend the most on housing suggests that women value comfortable living spaces more than other goods. The low expenditure on food for females suggests that females do not eat as much compared to males. Males spend the most on goods and the least on services. This suggests males prefer spending their incomes on manufactured items rather than serviced experiences. The fact that only the relationship between housing and service was statistically significant suggests that only service may have a true effect on housing that can be inferred on the general population.

Suicides2 data discussion

The suicides2 data analysis has shown that there are outliers within the data frame as seen in figure 4. This suggests that the calculated statistics may not give a full picture of what is going on and may misrepresent the dispersion of the data for the age groups of A35.44,A45.54, and A65.74. The standard deviation and mean values for these age groups may also be impacted by the presence of the outliers. The positively skewed data as seen in figure 5 shows that most of the suicides2 data is clustered towards the left tail of the distribution. It also shows that the average number of suicides is more than the median number of suicides. The distribution of suicides of different age groups faceted by countries-as seen in figure figure 6-suggests that Hungary may have the lowest quality of life/standard of living and lowest GDP per capita out of all the countries included in the data frame. This is because factors such as low standard of living and GDP per capita are things that can bring citizens of a country unhappiness and difficulty which could ultimately lead them to decide to commit suicide. The fact that Spain has the lowest number of suicides may suggest that they have the best quality of life out of the countries included within the data frame. It could also suggest that they have good mental health resources available. Annual holiday leave in Spain is also very generous, this may also help deter suicides. The statistical significance and R-squared value seen in the suicides2 linear regression summary shows that factors such as country and age can all truly affect the probability of an individual committing suicide.

Conclusion

Houshold data conclusion

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 for men and women (Hothorn & Everitt, 2014 [1]). The exploration and analysis of the data has shown that there is a clear relationship between total expenditure and expenditure between the four commodity groups. As total expenditure increases, so does spending across the commodity groups. This is the case for both males and females. The fact that males had a higher total expenditure than females and also a higher expenditure across the four commodity groups further makes this evident.

Suicides2 data conclusion

The objective of the suicides2 data frame was to construct different graphs such as box plots for the data from different age groups, and comment on what the graphics tell us about the data (Hothorn & Everitt, 2014 [1]). The box plots helped identify the presence of outliers within the data frame and gave an understanding of the standard deviation of the data. The violin plots gave an understanding of the relationship between suicides and age group and shed a light on the distribution of the data. The bar graph highlighted the difference between suicides within countries, and showed how suicides for different age groups differ depending on citizenship.

Appendices

```
data("household", package = "HSAUR3")
 data("suicides2", package = "HSAUR3")
 knitr::kable(household, "latex",caption = 'Household Data')
 knitr::kable(suicides2, "latex",caption = 'Suicides Data')
 summary(household)
 hl<-lm(housing~food+goods+service,data = household)
 summary(hl)
 nh<-household %>%
 mutate(total_expenditure=housing+food+goods+service)
 type_of_expenditure <- household %>%
        gather(Type_of_Expenditure,Expenditure, - c(gender))
 par(mfrow = c(1,4))
library(gridExtra)
plot1<-ggplot(nh, aes(x=housing, y=total_expenditure)) +
  geom_point(color="green")+geom_smooth(method=lm)+ theme(axis.text=element_text(size=6))
 plot2<-ggplot(nh, aes(x=food, y=total_expenditure)) +
  geom_point(color="red")+
  geom_smooth(method=lm)+ theme(axis.text=element_text(size=6))
 plot3<-ggplot(nh, aes(x=goods, y=total_expenditure))+geom_point(color="blue")+geom_smooth(method=lm)+ t
 plot4<-ggplot(nh, aes(x=service, y=total_expenditure)) +geom_point(color="yellow")+geom_smooth(method =
  grid.arrange(plot1, plot2, plot3,plot4, ncol=4)
```

```
## `geom_smooth()` using formula 'y ~ x'
## `geom_smooth()` using formula 'y ~ x'
## `geom_smooth()` using formula 'y ~ x'
## `geom_smooth()` using formula 'y ~ x'
```

```
 p<-ggplot(type_of_expenditure, aes(x = Type_of_Expenditure, y = Expenditure)) +
  geom_col()
  p+facet_grid(gender~ .)
```

```
ggplot(type_of_expenditure, aes(x=Expenditure, color=Type_of_Expenditure)) +
 geom_density()+facet_grid(gender~ .)
```

```
suicides<-suicides2
suicides <- suicides %>%
       pivot_longer(cols= c(A25.34,A35.44,A45.54,A55.64,A65.74),names_to = "Age_Group", values_to = "Su
summary(suicides2)
suicides2 <- cbind(Countries = rownames(suicides2), suicides2)

rownames(suicides2) <- NULL

suicides3 <- suicides2 %>%
       gather(Age_Group,Suicides, - c(Countries))

s4<-lm(Suicides~Countries+Age_Group,data = suicides3)
summary(s4)
ggplot(suicides, aes(x=Age_Group, y=Suicides)) +
   geom_boxplot()
```

```
dp <- ggplot(suicides3, aes(x=Age_Group, y=Suicides, fill=Age_Group)) +
 geom_violin(trim=FALSE)+
 geom_boxplot(width=0.1, fill="white")+
 labs(title="Plot of Suicides  by Age Group",x="Age Group", y = "Suicides")

dp + theme_classic()
```

```
ggplot(suicides3, aes(fill=Countries, y=Suicides, x=Age_Group)) +
   geom_bar(position="stack", stat="identity")
```

References

[1] Hothorn, T and Everitt, B. (2014) *A Handbook of Statistical Analyses Using R* Florida: CRC Press.