Assignment_WK5_Saripalli

Balaram

2025-02-17

Assignment Week 5 - ANOVA

1. Form a hypothesis for variables that maybe related.
2. Write a null and alternative hypothesis
3. Create a boxplot of your variables.
4. Run ANOVA on variables in your hypothesis.
5. Run a post hoc test to measure significant between factors
6. Is there a significant interaction effect between the levels of each variable? Create at least one interaction plot.
7. Test for ANOVA assumptions.
8. Does the analysis support the hypothesis you formed initially?
9. Post your Rmd file and knitted pdf file to the assignment dropbox.

Marketing Data Analysis

Setting the working directory

```
setwd("F:/Balaram/Statcourse")
```

Loading the required Packages

```
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v readr     2.1.5
## v forcats   1.0.0     v stringr   1.5.1
## v ggplot2   3.5.1     v tibble    3.2.1
## v lubridate 1.9.4     v tidyr     1.3.1
## v purrr     1.0.2
## -- Conflicts ---------------------------------------- tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```
library(car)
```

```
## Loading required package: carData
##
```

```
## Attaching package: 'car'
##
## The following object is masked from 'package:dplyr':
##
##     recode
##
## The following object is masked from 'package:purrr':
##
##     some
```

```
library(ggplot2)
```

1. Form a hypothesis for variables that maybe related.

Based on the marketing dataset, We want to know weather the education levels, marital statuses, and age groups significantly affect the total amount spent on wine purchases.

Hypothesis formulation: effect of education level, marital staus and age groups on the total amount spend

Factors considered: education levels, marital statuses, and age groups. lets considered the effect of these factors on total amount spend.

2. Write a null and alternative hypothesis

Null Hypothesis (H0): There is no significant difference in total amount spent across different education levels, marital statuses, and age groups.

Alternative Hypothesis (H1): There is a significant difference in total amount spent across different education levels, marital statuses, and age groups.

Read the data

```
marketing_data <- read.csv("marketing.csv")
View(marketing_data)
```

Create total amount spent variable

As we are focusing on the total amount spent and that variable is not present in the data we are creating that variable which is equal to the sum of all the amounts spent on different products

```
marketing_data$TotalAmountSpent <- marketing_data$MntWines +
                                   marketing_data$MntFruits +
                                   marketing_data$MntMeatProducts +
                                   marketing_data$MntFishProducts +
                                   marketing_data$MntSweetProducts +
                                   marketing_data$MntGoldProds
```

Create age groups

We are also creating a variable age group using the maximum value of Date of customer in the data (2013) - the year of birth and we break it in to 4 classes

```
marketing_data$AgeGroup <- cut(2013 - marketing_data$Year_Birth,
                               breaks = c(0, 30, 45, 60, Inf),
                               labels = c("18-30", "31-45", "46-60", "60+"))
```

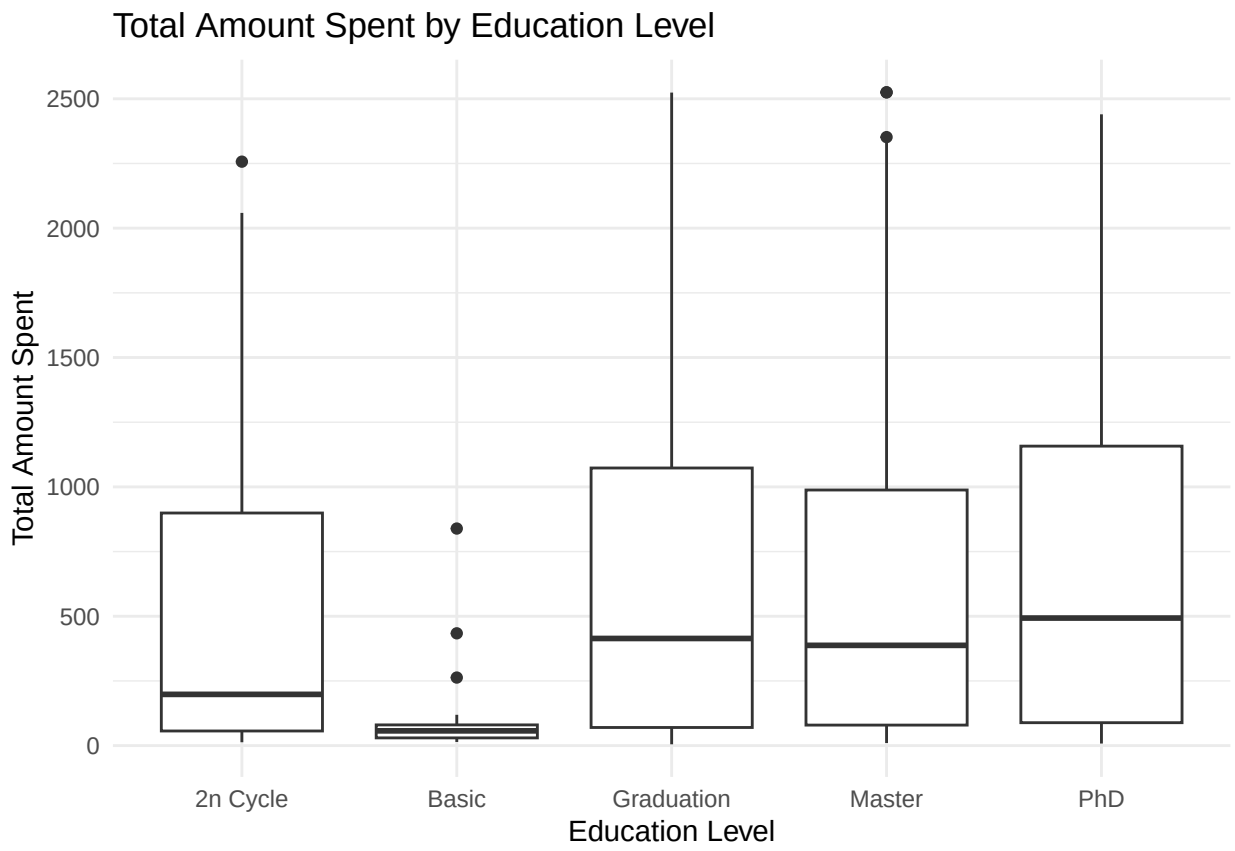
Data preprocessing i.e., remove the missing values if any

```
# Remove rows with missing values
marketing_data <- na.omit(marketing_data)
```

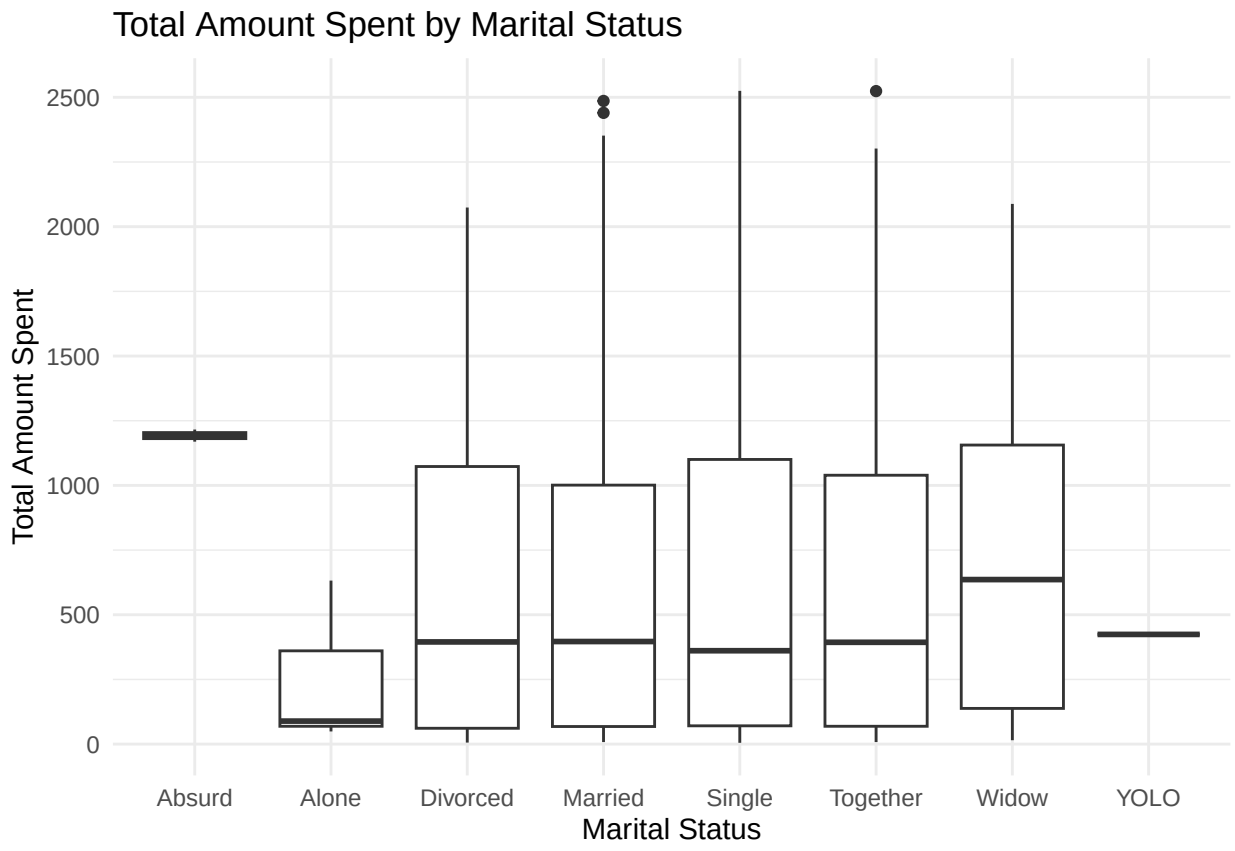
3. Create a boxplot of your variables.

Boxplots are created to visualize the relationship between our variables:

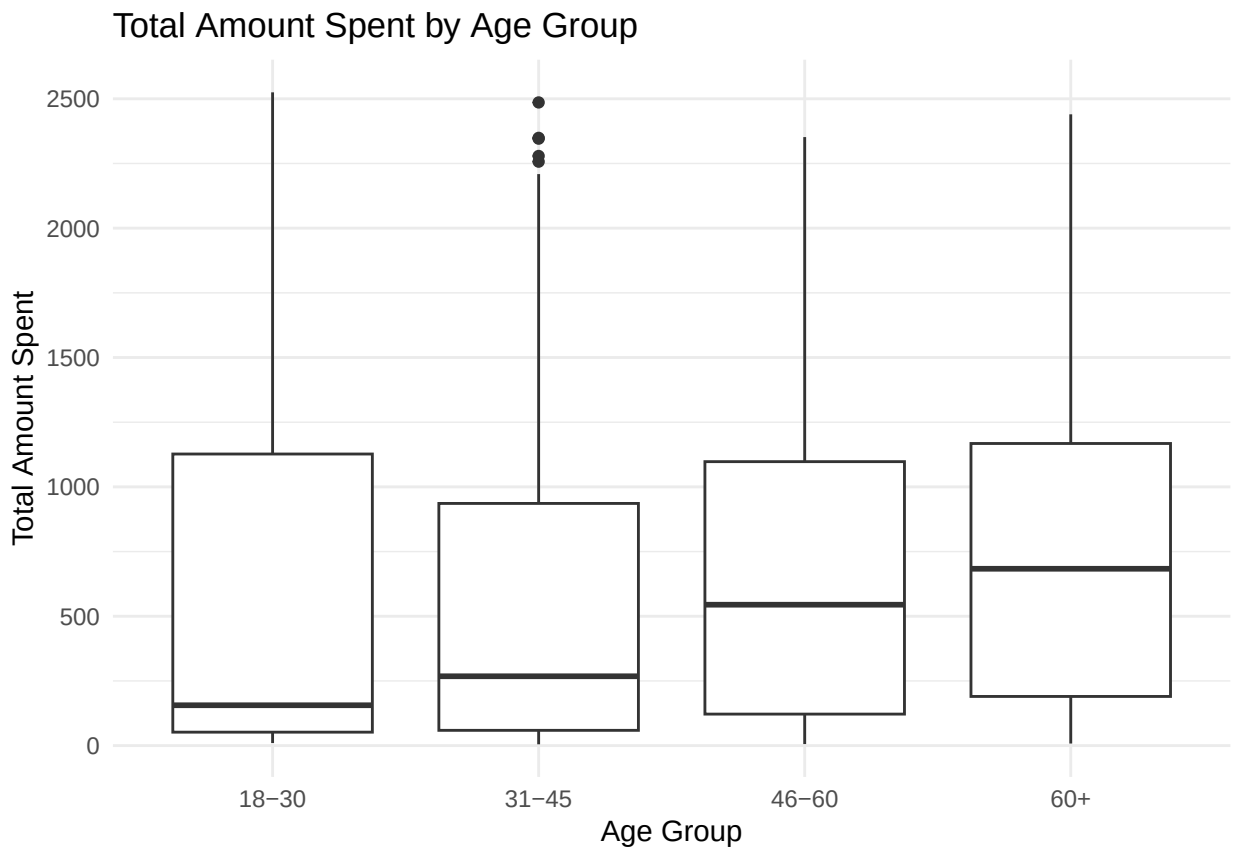
```
# Boxplot for Education
ggplot(marketing_data, aes(x = Education, y = TotalAmountSpent)) +
  geom_boxplot() +
  theme_minimal() +
  labs(title = "Total Amount Spent by Education Level",
       x = "Education Level",
       y = "Total Amount Spent")
```



```r
# Boxplot for Marital Status
ggplot(marketing_data, aes(x = Marital_Status, y = TotalAmountSpent)) +
  geom_boxplot() +
  theme_minimal() +
  labs(title = "Total Amount Spent by Marital Status",
       x = "Marital Status",
       y = "Total Amount Spent")
```




```r
# Boxplot for Age Group
ggplot(marketing_data, aes(x = AgeGroup, y = TotalAmountSpent)) +
  geom_boxplot() +
  theme_minimal() +
  labs(title = "Total Amount Spent by Age Group",
       x = "Age Group",
       y = "Total Amount Spent")
```



Analysis of box plots Box plots were created to visualize the relationship between total amount spent and the variables of interest:

1. Education: - Spending increases with education level. Customers with "Basic" education spend the least, while those with "PhD" spend the most.

- Significant differences were observed between groups (e.g., "Basic" vs. "PhD"), supporting the hypothesis that education affects total spending.

2. Marital Status: - Spending differences across marital statuses were minimal and not statistically significant.

- This suggests that marital status does not strongly influence total spending.

Age Group: - Older age groups (46–60 and 60+) tend to spend more than younger groups (18–30 and 31–45).

- Significant differences support the hypothesis that age influences spending.

4. Run ANOVA on variables in your hypothesis.

```
anova_model <- aov(TotalAmountSpent ~ Education + Marital_Status + AgeGroup, data = marketing_data)
summary(anova_model)
```

```
##                   Df    Sum Sq Mean Sq F value   Pr(>F)
## Education          4  19644802 4911201  14.000 2.77e-11 ***
## Marital_Status     7   2551305  364472   1.039    0.401
## AgeGroup           3   9386370 3128790   8.919 7.13e-06 ***
## Residuals       2225 780512237  350792
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

As we have done one- way ANOVA. we will get that weather these factors have significant effect on the total amount spent.

Interpretation The ANOVA results show which factors are statistically significant. Look for p-values $<$ 0.05 to identify significant factors.

Based on the 'p-values' the factors education and age group are statistically significant and marital status was not significant.

5. Run a post hoc test to measure significant between factors

If ANOVA shows significant differences, we'll perform a Tukey HSD test:

```
tukey_results <- TukeyHSD(anova_model)
print(tukey_results)
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = TotalAmountSpent ~ Education + Marital_Status + AgeGroup, data = marketing_data)
##
## $Education
##                             diff           lwr        upr     p adj
## Basic-2n Cycle      -414.730797 -662.30840589 -167.15319 0.0000497
## Graduation-2n Cycle  123.371753    0.08808572  246.65542 0.0497335
## Master-2n Cycle      115.253987  -25.97314165  256.48112 0.1697162
## PhD-2n Cycle         175.882371   40.75816442  311.00658 0.0035596
## Graduation-Basic     538.102550  312.85721179  763.34789 0.0000000
## Master-Basic         529.984785  294.43924379  765.53033 0.0000000
## PhD-Basic            590.613169  358.67535120  822.55099 0.0000000
## Master-Graduation     -8.117765 -104.99857983   88.76305 0.9993941
## PhD-Graduation        52.510619  -35.23527321  140.25651 0.4758340
## PhD-Master            60.628384  -50.93137998  172.18815 0.5732506
##
## $Marital_Status
##                           diff         lwr       upr     p adj
## Alone-Absurd      -954.689834 -2594.97075  685.5911 0.6435065
## Divorced-Absurd   -581.854322 -1857.87524  694.1666 0.8650389
## Married-Absurd    -592.107059 -1864.13290  679.9188 0.8519060
## Single-Absurd     -569.839315 -1843.03969  703.3611 0.8761126
## Together-Absurd   -572.166872 -1844.91174  700.5780 0.8735489
## Widow-Absurd      -457.843043 -1744.79413  829.1080 0.9611463
## YOLO-Absurd       -825.069501 -2621.90722  971.7682 0.8607063
## Divorced-Alone     372.835511  -671.25504 1416.9261 0.9603503
## Married-Alone      362.582775  -676.62145 1401.7870 0.9650579
```

```
## Single-Alone       384.850519  -655.79106 1425.4921 0.9521548
## Together-Alone     382.522962  -657.56126 1422.6072 0.9535420
## Widow-Alone        496.846791  -560.57403 1554.2676 0.8456098
## YOLO-Alone         129.620333 -1510.66059 1769.9013 0.9999977
## Married-Divorced   -10.252737  -143.11856  122.6131 0.9999981
## Single-Divorced     12.015007  -131.66100  155.6910 0.9999967
## Together-Divorced    9.687451  -129.89438  149.2693 0.9999991
## Widow-Divorced     124.011279  -112.30778  360.3303 0.7553012
## YOLO-Divorced     -243.215179 -1519.23610 1032.8057 0.9991230
## Single-Married      22.267744   -80.02174  124.5572 0.9979323
## Together-Married    19.940187   -76.51411  116.3945 0.9985153
## Widow-Married      134.264016   -79.43446  347.9625 0.5467892
## YOLO-Married      -232.962442 -1504.98828 1039.0634 0.9993246
## Together-Single     -2.327557  -113.20091  108.5458 1.0000000
## Widow-Single       111.996272  -108.58593  332.5785 0.7854429
## YOLO-Single       -255.230186 -1528.43056 1017.9702 0.9987829
## Widow-Together     114.323829  -103.61378  332.2614 0.7556482
## YOLO-Together     -252.902630 -1525.64749 1019.8422 0.9988504
## YOLO-Widow        -367.226458 -1654.17754  919.7246 0.9889698
##
## $AgeGroup
##                   diff         lwr       upr     p adj
## 31-45-18-30 -53.51675 -153.803057  46.76957 0.5171298
## 46-60-18-30  61.58479  -42.931967 166.10154 0.4286449
## 60+-18-30   125.46865   -7.242761 258.18005 0.0717018
## 46-60-31-45 115.10153   40.439403 189.76366 0.0004423
## 60+-31-45   178.98539   68.246965 289.72382 0.0001973
## 60+-46-60    63.88386  -50.699764 178.46748 0.4785346
```

Based on the Tukey multiple comparisons of means results, we can interpret the post-hoc analysis as follows:

Education There are significant differences in total amount spent between several education levels:

1. Basic education level spends significantly less than all other education levels ($p < 0.05$ for all comparisons with Basic).
2. 2n Cycle education level spends significantly less than Graduation ($p = 0.0497$) and PhD ($p = 0.0036$).
3. There are no significant differences between Graduation, Master, and PhD education levels.

Marital Status

1. No statistically significant differences were found between marital status groups in terms of total amount spent (all p-values > 0.05).
2. The large p-values and wide confidence intervals suggest high variability within groups or small sample sizes for some categories (e.g., Absurd, Alone, YOLO).

Age Group 1. There are significant differences between some age groups:

2. The 46-60 age group spends significantly more than the 31-45 age group ($p = 0.0004$).
3. The 60+ age group spends significantly more than the 31-45 age group ($p = 0.0002$).

4. There is a marginally significant difference between the 60+ and 18-30 age groups (p = 0.0717), with the 60+ group tending to spend more.

5. No significant differences were found between other age group comparisons.

Overall Interpretation 1. Education level appears to have the most consistent impact on total amount spent, with Basic education level associated with significantly lower spending.

2. Marital status does not seem to have a significant effect on total amount spent, though this could be due to high variability or small sample sizes in some categories.

3. Age has some influence on spending, with older age groups (46-60 and 60+) tending to spend more than the 31-45 age group.

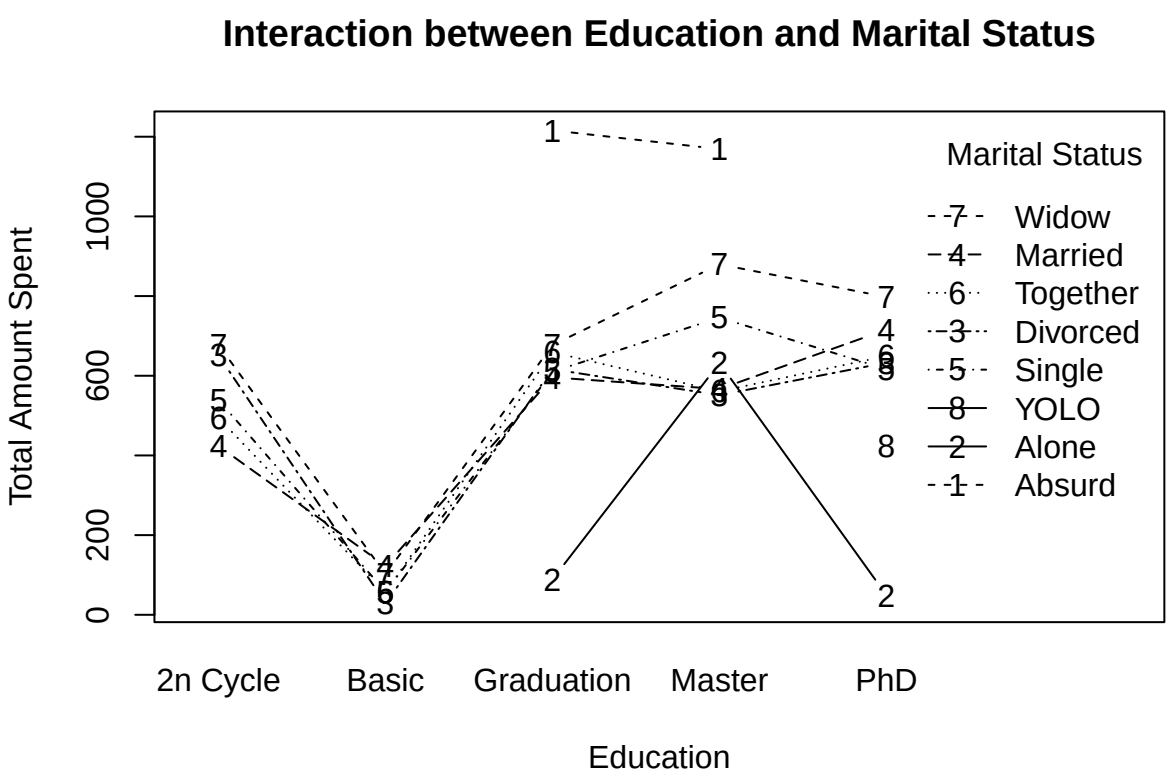
These results suggest that marketing strategies might be most effectively tailored based on education level and age group rather than marital status.This is also in allignment with the one-way ANOVA analysis conducted.

6. Is there a significant interaction effect between the levels of each variable? Create at least one interaction plot.

```
interaction_model <- aov(TotalAmountSpent ~ Education * Marital_Status, data = marketing_data)
summary(interaction_model)
```

```
##                            Df    Sum Sq Mean Sq F value   Pr(>F)
## Education                   4  19644802 4911201  13.824 3.87e-11 ***
## Marital_Status              7   2551305  364472   1.026    0.411
## Education:Marital_Status   19   5093808  268095   0.755    0.763
## Residuals                2209 784804800  355276
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
# Interaction plot
interaction.plot(marketing_data$Education, marketing_data$Marital_Status,
                 marketing_data$TotalAmountSpent,
                 type = "b",
                 xlab = "Education",
                 ylab = "Total Amount Spent",
                 trace.label = "Marital Status",
                 main = "Interaction between Education and Marital Status")
```



Interpretation:

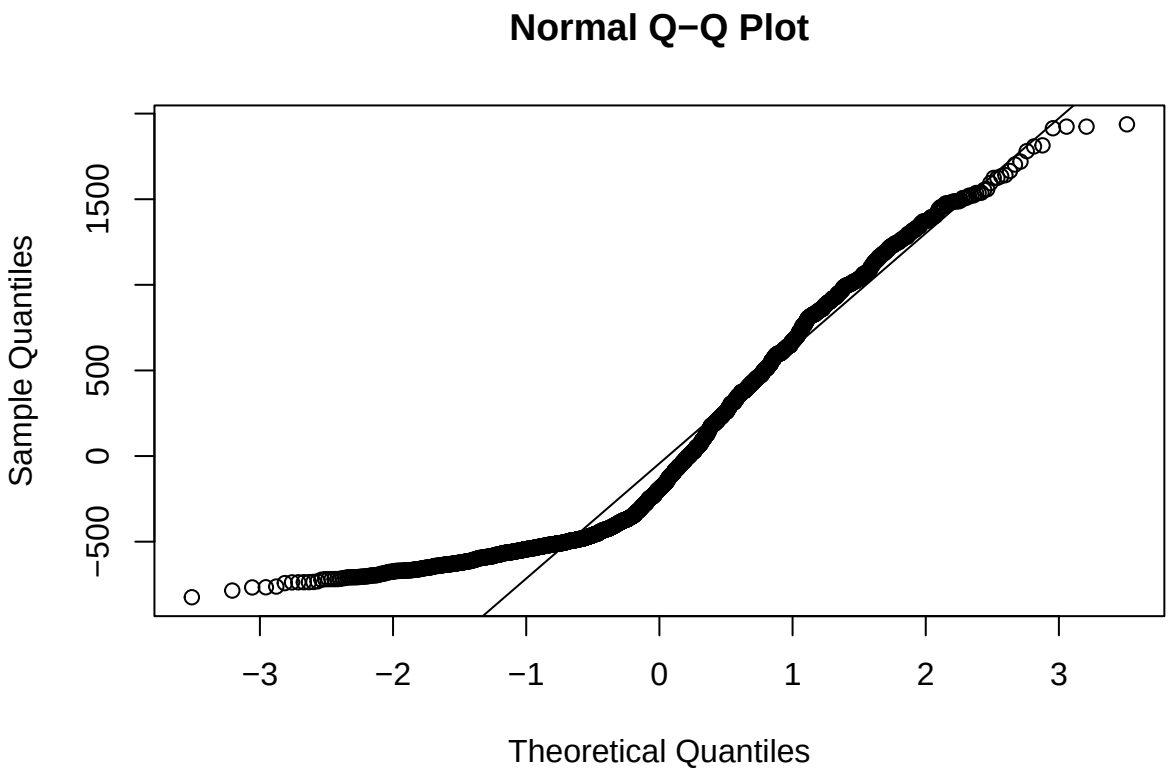
The interaction plot visually suggests some variability in spending patterns across education levels and marital statuses, with "Basic" education consistently associated with the lowest spending and higher education levels (e.g., PhD) showing increased spending. This suggested that there is interaction effects.

However, the p-value of 0.763 for the interaction effect indicates that these differences are not statistically significant. This means that the combined effect of education and marital status on total amount spent is likely due to random variation rather than a meaningful interaction.

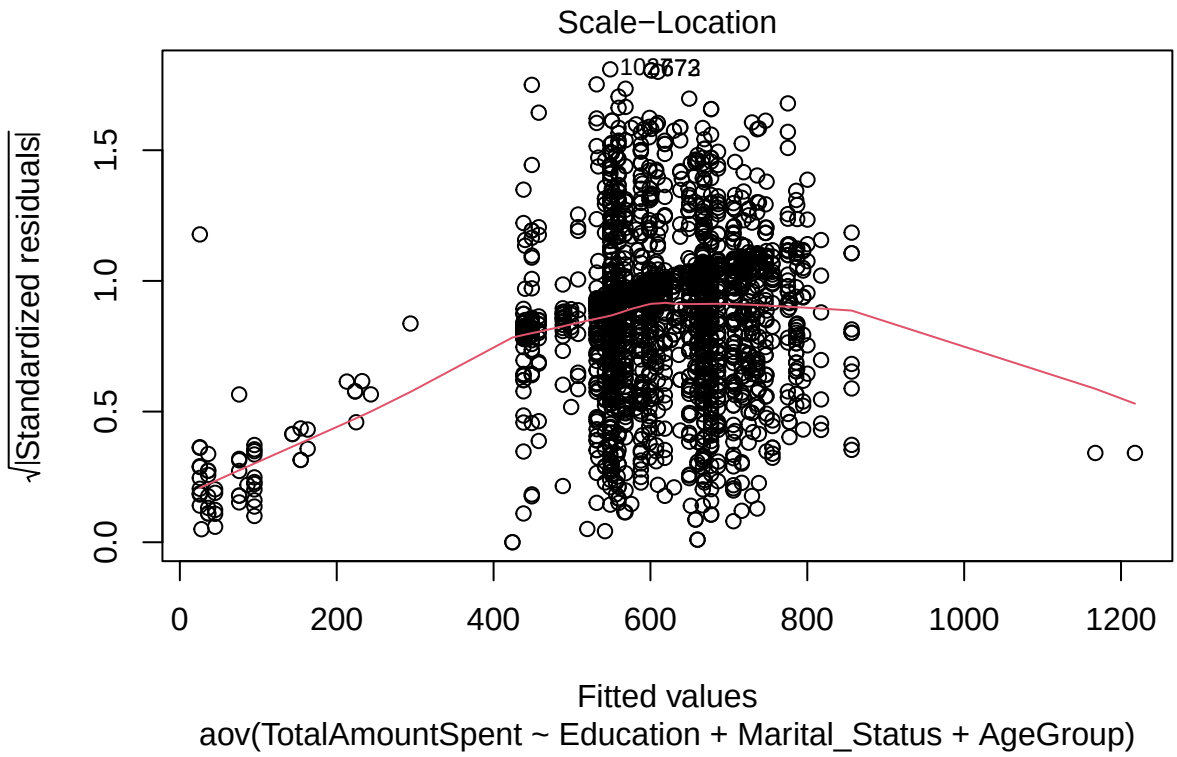
7. Test for ANOVA assumptions.

ANOVA involves three major assumptions i.e., normality, homegenity of variances and levens test

```
# Normality test
qqnorm(residuals(anova_model))
qqline(residuals(anova_model))
```




```r
# Homogeneity of variances
plot(anova_model, which = 3)
```




```r
# Levene's test
leveneTest(TotalAmountSpent ~ Education * Marital_Status * AgeGroup, data = marketing_data)
```

```
## Levene's Test for Homogeneity of Variance (center = median)
##         Df F value    Pr(>F)
## group   94  1.7188 3.176e-05 ***
##       2145
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Normal Q-Q Plot The residuals mostly follow the diagonal line, indicating approximate normality.Deviations at the tails suggest some outliers or non-normality in extreme values.

Scale-Location Plot The red trend line is slightly curved, indicating mild heteroscedasticity (non-constant variance).Variance appears to increase slightly with higher fitted values, but the deviation is not severe. Based on the Levene's test results provided, the p-value is $< 2.2e\text{-}16$, which is less than 0.05. This indicates that the test is statistically significant.

Levens test The null hypothesis of Levene's test is that the variances are equal across groups.With a p-value < 0.05, we reject the null hypothesis. This means there is strong evidence that the variances are not equal across groups (i.e., there is heterogeneity of variances).

As p is in the order of e-5, i,e, < 0.05 we can reject the null hypothesis thus, there is a significant variation between the variances across the groups.

Conclusion The assumptions of normality and homoscedasticity are slightly violated but not critically, so ANOVA results can still be interpreted with caution.

8. Does the analysis support the hypothesis you formed initially?

Based on the ANOVA results, we can determine whether to reject or fail to reject the null hypothesis. We'll interpret the p-values and effect sizes to understand the significance and magnitude of the relationships between our variables and total amount spent.

The initial hypothesis was that education level, marital status, and age group affect total spending:

1. Education: Supported – Education has a significant impact on total spending.
2. Marital Status: Not Supported – Marital status does not significantly influence spending. (As $p > 0.05$)
3. Age Group: Partially Supported – Age group significantly affects spending, but differences are more pronounced between specific groups.

Conclusions

- Education level and age group are significant factors influencing total spending, with higher education levels and older age groups associated with greater spending.
- Marital status does not significantly impact total spending.