MSDS660_Week5_Assignment_Apeetz

2022-11-13

Adam Peetz

MSDS660 Week 5 Discussion

Regis University

Dr. Siripun Sanguansintukul

November 20th 2022

ANOVA on Marketing Data

1. Form a hypothesis for the variables that maybe related. You may have both factors and numerical values in your analysis. You would need factors to create an interaction plot.

Hypothesis: NumCatalogPurchases and Education will significantly impact MntMeatProducts received with a significant interaction between them indicating that the levels of one variable will affect the levels of another variable and will vary depending on the categories.

Data: The data used in this notebook is marketing data provided by the Hult International School of Business. All observation are independent; each row in this data set corresponds to a unique customer.

```
# load libraries
library(tidyverse)
library(data.table)
library(ggpubr)
library(car)
library(dplyr)
library(agricolae)
library(rstatix)

# load data
data <- read_csv("marketing.csv",show_col_types = FALSE)
# convert data to table
df<-as.data.table(data)
```

Check for NAs, Treat Outliers, and Feature Engineering

NAs: There are no NAs in the dataset.

Outliers: The features of MntMeatProducts and NumCatalogPurchases both contain outliers. Most customers have made fewer than 10 catalog purchases for less than 500 dollars' worth of meat. A few customers in the dataset have made more than twenty catalog purchases and spent more than 1,000 dollars on meat. These outliers are removed from the dataset using a quantile clipping method in the code below.

Feature Engineering: Research into the labels in the education feature revealed “2n Cycle” refers to students who have completed graduate degrees. This label is the same as “Master”, and may be a result of regional differences in this international dataset. “2n Cycle” and “Master” will be combined into a single category. All education labels are also changed to United States style labels for High School education, Bachelors, Masters, and Doctors.

Removing High School: High school or “basic” educated customers make up around 2% of the customers in the database. This is a small sample size when compared to the groups for Bachelors, Masters, and Doctorate educated customers. Rows corresponding to High School educated customers will be removed from the dataset.

```
# check for NAs in education
sum(is.na(df$Education))
```

```
## [1] 0
```

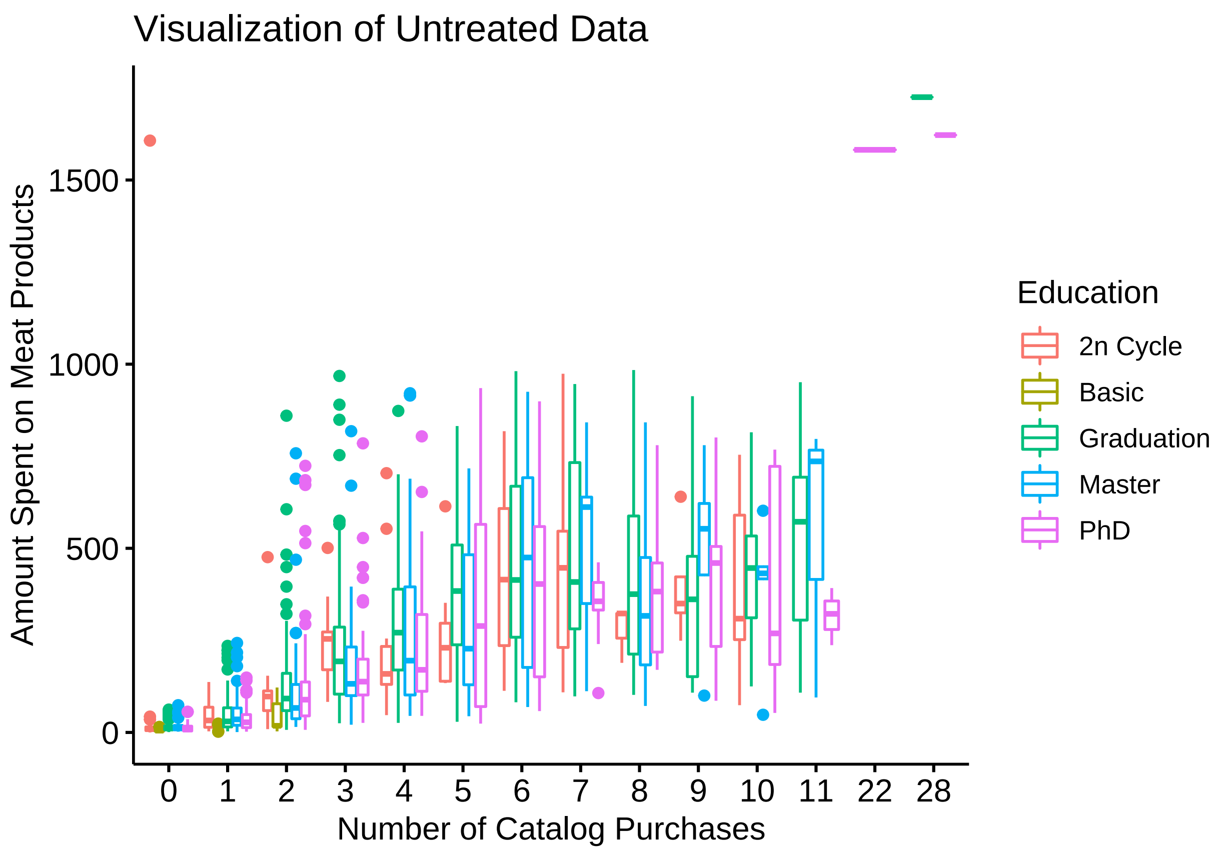
```
# check for NAs in MntMeatProducts
sum(is.na(df$MntMeatProducts))
```

```
## [1] 0
```

```
# check for NAs in NumCatalogPurchases
sum(is.na(df$NumCatalogPurchases))
```

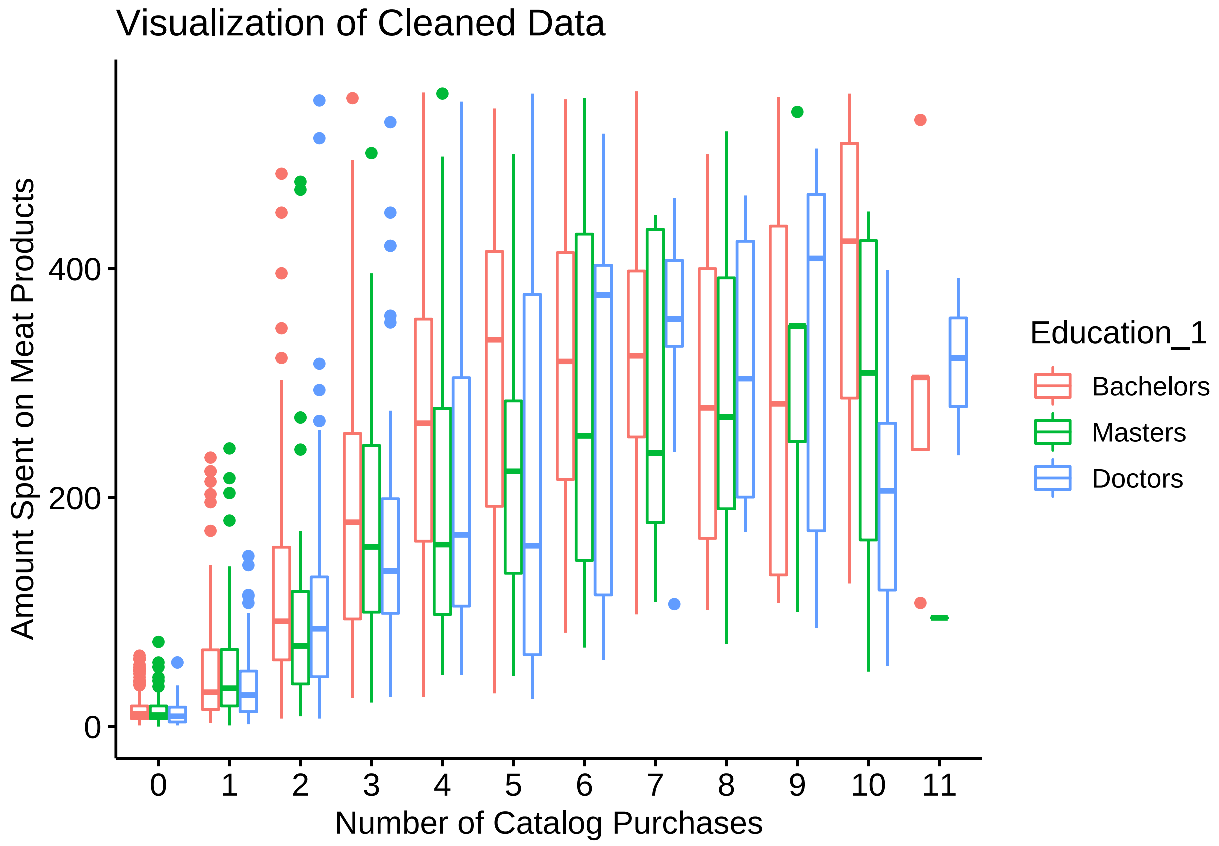
```
## [1] 0
```

```
# check normality of distributions
ggboxplot(df, x = "NumCatalogPurchases", y = "MntMeatProducts", color = "Education",
          main = "Visualization of Untreated Data",
          xlab = "Number of Catalog Purchases",
          ylab = "Amount Spent on Meat Products") +
          theme(legend.position="right")
```



```
# combine 2nd Cycle and Masters
df<-df %>% mutate(Education_1 = case_when(Education == "PhD" ~ "Doctors",
                                        Education == "Master" ~ "Masters",
                                        Education == "2n Cycle" ~ "Masters",
                                        Education == "Graduation" ~ "Bachelors",
                                        Education == "Basic" ~ "High School") %>%
           fct_relevel("High School",
                       "Bachelors",
                       "Masters",
                       "Doctors"))

# Treat outliers
# calculate quantiles and interqauntile range
Q <- quantile(df$MntMeatProducts, probs=c(.25, .75), na.rm = TRUE)
iqr <- IQR(df$MntMeatProducts, na.rm = TRUE)

# create upper and lower quantile ranges
up <-  Q[2]+1.5*iqr # Upper Range
low<- Q[1]-1.5*iqr # Lower Range

#remove outliers with quantile data
cleaned_df_1 <- subset(df, df$MntMeatProducts > (Q[1] - 1.5*iqr) & df$MntMeatProducts < (Q[2]+1.5*iqr))

#subset to single engine planes
cleaned_df_1 = cleaned_df_1[cleaned_df_1$Education_1  %in% c("Doctors", "Masters", "Bachelors"), ]
```

```
#display treatment results
ggboxplot(cleaned_df_1, x = "NumCatalogPurchases", y = "MntMeatProducts", color = "Education_1",
          main = "Visualization of Cleaned Data",
          xlab = "Number of Catalog Purchases",
          ylab = "Amount Spent on Meat Products") +
          theme(legend.position="right")
```



Multiway ANOVA

2. Run a multi-way ANOVA on loan amount received with at least 2 other variables.

Ho: The mean outcome is the same across all groups.

Ha: At least one mean is different.

Results P-values less than 0.05 for NumCatalogPurchases and Education allow the null hypothesis to be rejected for those variables, there is a significant difference in the mean outcome for these groups.

The p-value for the interaction between total_credit_limit and application_type is 0.002, which indicates that the relationships between Education and MntMeatProducts depend on the NumCatalogPurchases.

```
multi_way_model<-aov(MntMeatProducts ~ NumCatalogPurchases * Education_1, data=cleaned_df_1)
summary(multi_way_model)
```

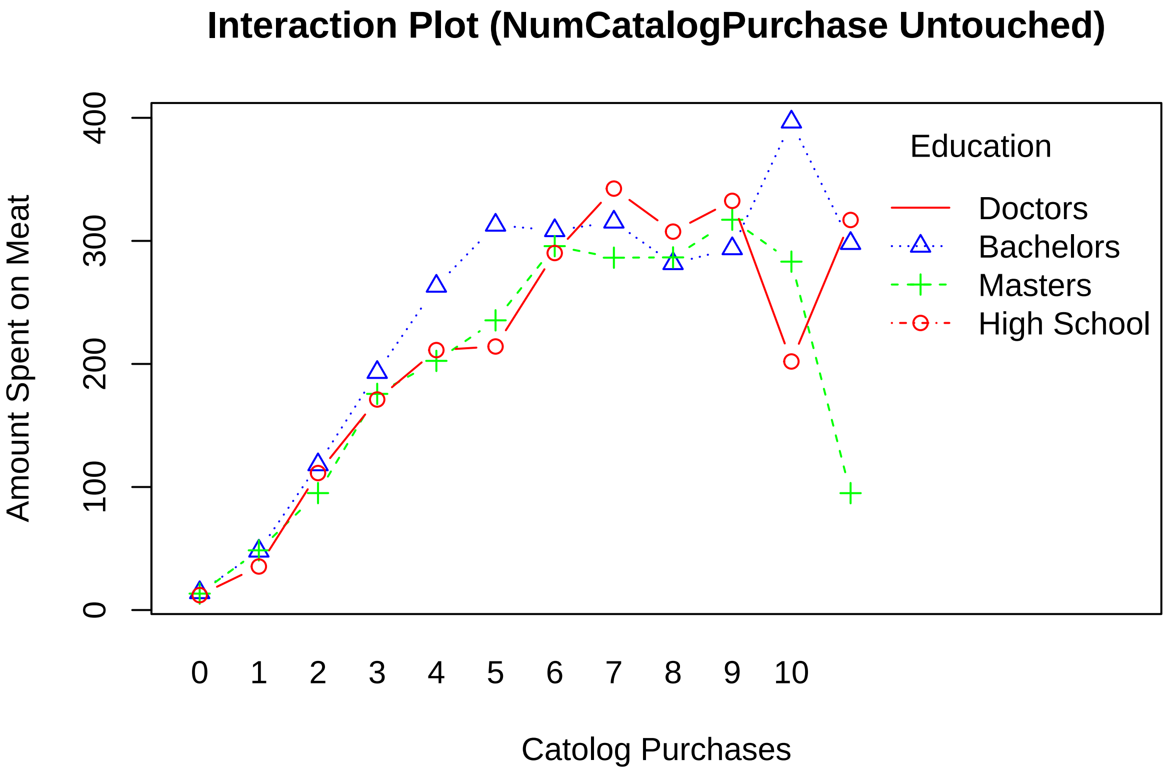
```
##                                    Df   Sum Sq  Mean Sq  F value    Pr(>F)
## NumCatalogPurchases                 1 21040848 21040848 2280.837  < 2e-16 ***
## Education_1                         2   154172    77086    8.356 0.000243 ***
## NumCatalogPurchases:Education_1    2   113998    56999    6.179 0.002113 **
## Residuals                       2005 18496234     9225
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Interaction Plot

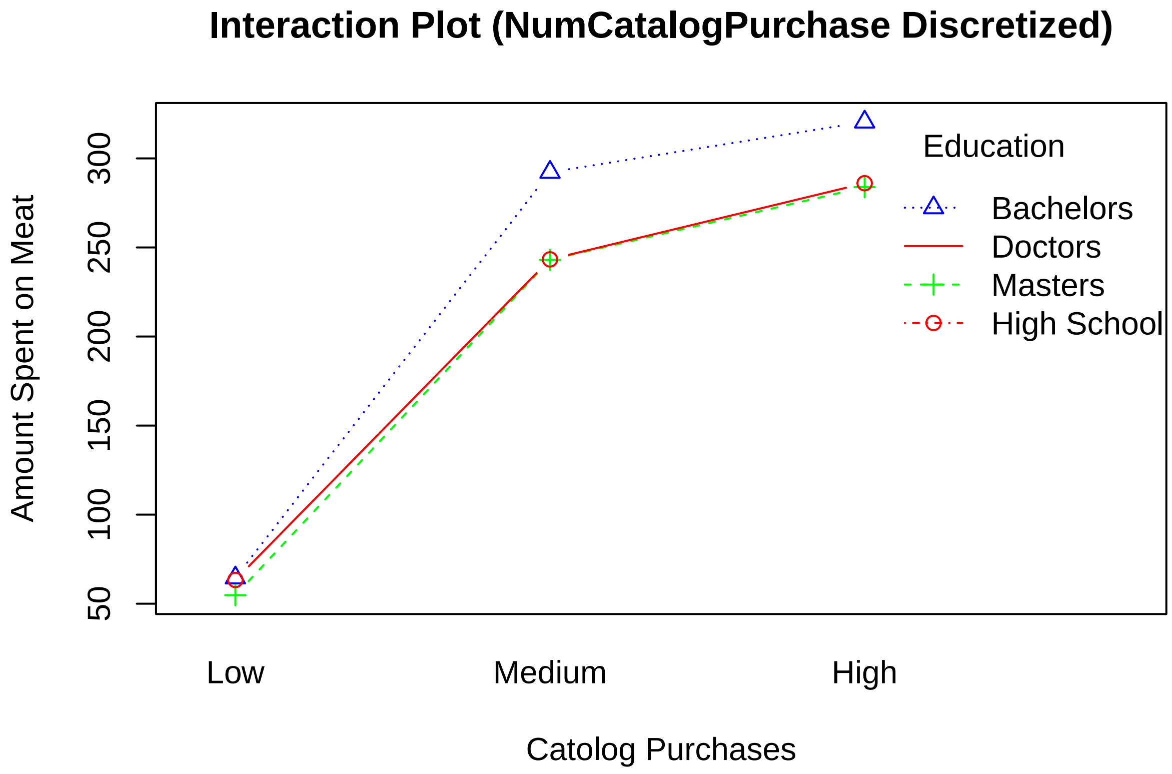
3. Is there a significant interaction effect between the levels of each variable? Please plot at least one interaction plot.

The interaction plots for these variables either show or hide interaction depending on how the NumCatalogPurchases feature is treated. Combining NumCatalogPurchases into a 3-category feature for "Low", "Middle", and "High" amounts of catalog purchases shows no interaction between the variables at different levels of catalog purchases. Leaving NumCatalogPurchases untouched shows interaction for the highest levels of catalog purchases.

```r
cleaned_df_1 <- within(cleaned_df_1, {
  catolog_cat <- NA # need to initialize variable
  catolog_cat[NumCatalogPurchases <= 3] <- "Low"
  catolog_cat[NumCatalogPurchases >= 4 & NumCatalogPurchases <= 7] <- "Medium"
  catolog_cat[NumCatalogPurchases >= 8] <- "High"
   } )

cleaned_df_1$catolog_cat <- factor(cleaned_df_1$catolog_cat, levels = c("Low", "Medium", "High"))

interaction.plot(x.factor = cleaned_df_1$NumCatalogPurchases,
                 trace.factor = cleaned_df_1$Education_1,
                 response = cleaned_df_1$MntMeatProducts,
                 fun = mean,
                 type = "b",  # shows each point
                 main = "Interaction Plot (NumCatalogPurchase Untouched)",
                 legend = TRUE,
                 trace.label = "Education",
                 xlab = "Catolog Purchases",
                 ylab="Amount Spent on Meat",
                 pch=c(1, 2, 3),
                 col = c("Red", "Blue", "Green"))
```



```
interaction.plot(x.factor = cleaned_df_1$catolog_cat,
                 trace.factor = cleaned_df_1$Education_1,
                 response = cleaned_df_1$MntMeatProducts,
                 fun = mean,
                 type = "b",  # shows each point
                 main = "Interaction Plot (NumCatalogPurchase Discretized)",
                 legend = TRUE,
                 trace.label = "Education",
                 xlab = "Catolog Purchases",
                 ylab="Amount Spent on Meat",
                 pch=c(1, 2, 3),
                 col = c("Red", "Blue", "Green"))
```



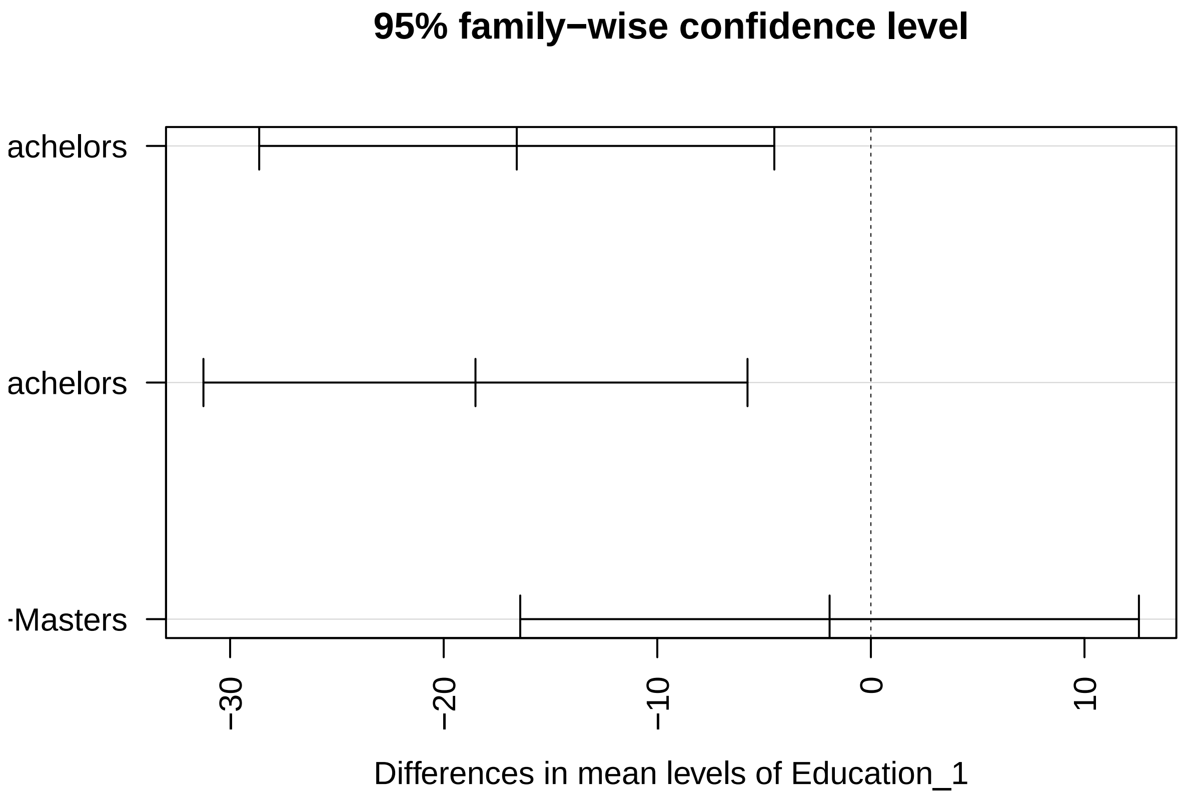
Tukey HSD

Insight into which group means are different can be gained using a Tukey Test. The Tukey Test shows that there are significant differences between means for the Bachelors group and the Masters/Doctors group but that the means for Masters and Doctors are not significantly different.

```
TukeyHSD(multi_way_model, which = "Education_1")
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = MntMeatProducts ~ NumCatalogPurchases * Education_1, data = cleaned_df_1)
##
## $Education_1
##                           diff       lwr       upr     p adj
## Masters-Bachelors -16.579400 -28.63828 -4.520516 0.0036659
## Doctors-Bachelors -18.512821 -31.24916 -5.776479 0.0019205
## Doctors-Masters    -1.933421 -16.41639 12.549544 0.9473963
```

```
plot(TukeyHSD(multi_way_model, which = "Education_1"), las=2)
```



Levene’s Test for Homogeneity of Variances

4. Test for ANOVA assumptions. (At least the Levene’s test for HOV)

To perform ANOVA, the data must meet a few assumptions. Levene’s test confirms the homogeneity of variance where f-values less than 0.05 indicate a violation of the homogeneity assumption.

Ho: All populations variances are equal.

Ha: At least two variances differ.

An p-value of less than 0.05 indicates that the null hypothesis should be rejected. At least two variances differ.

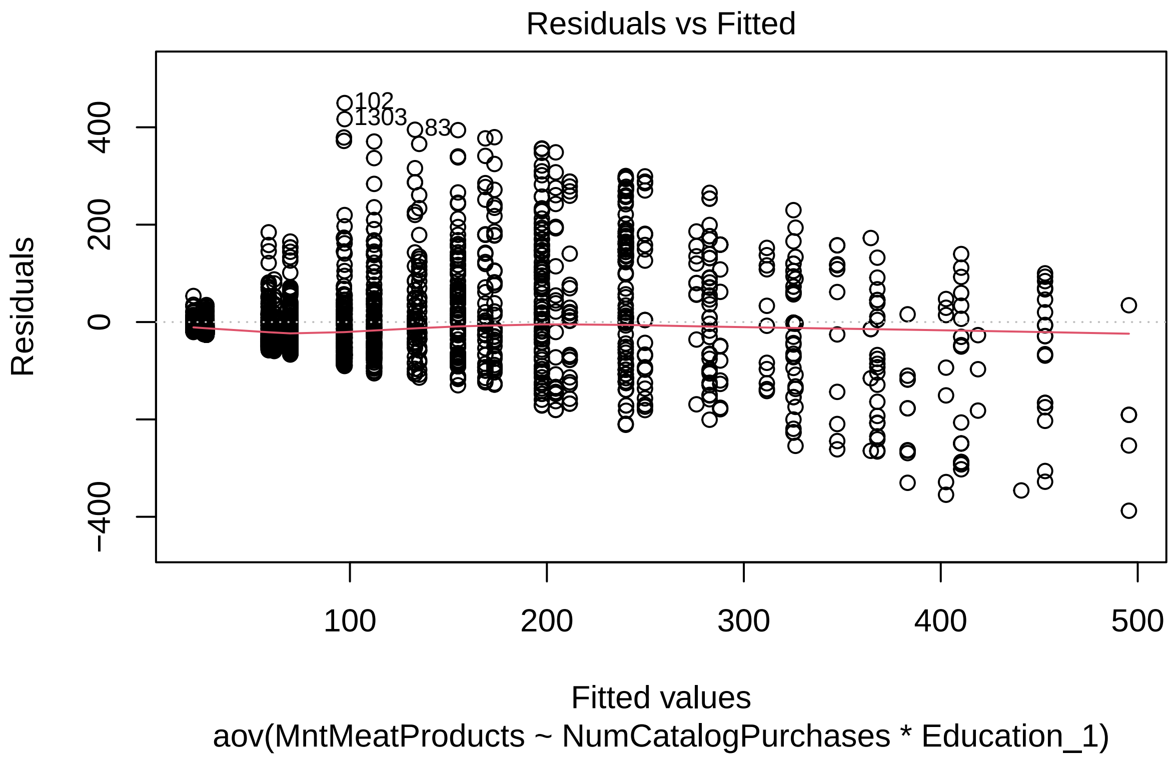
```r
# generate levene test
result = leveneTest(MntMeatProducts ~ interaction(Education_1, NumCatalogPurchases),
                    data = cleaned_df_1)

# display levene test results
print(result)
```

```
## Levene's Test for Homogeneity of Variance (center = median)
##         Df F value    Pr(>F)
## group   35  32.844 < 2.2e-16 ***
##       1975
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Homogeneity of Variances by Plotting Residuals vs Fitted

Homogeneity can also be proven by plotting residual vs fitted values. Data with homogeneous variance would show an even spread of values across the chart with similar density occurring above and below the center line. The chart for the selected marketing data starts with densely packed values on the left and spreads out with a downward trend as the fitted values increase. This chart further proves that the data violates the assumption of homogeneity.

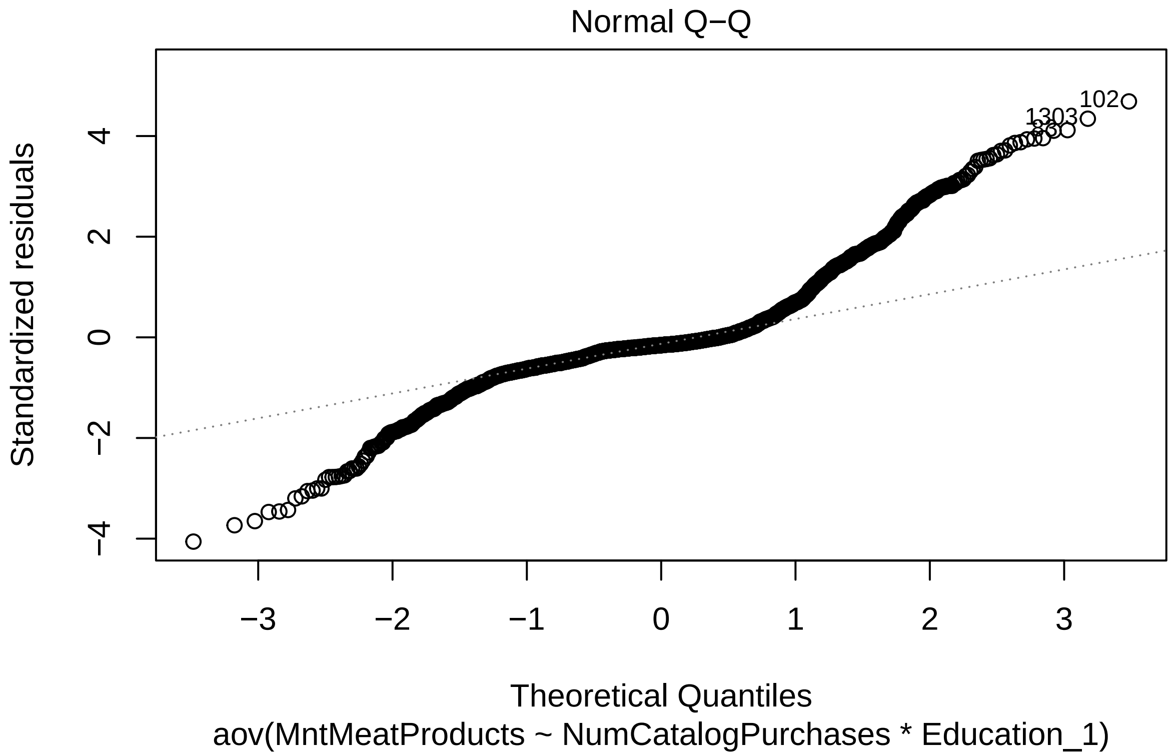
```
plot(multi_way_model, 1)
```



Normality

Normality can be proven with a Q-Q plot. Data that meets the assumption of normality should arrange itself along the dotted line of the Q-Q plot. The Q-Q plot for the selected markeing data shows departures from normality on either end of the plot. The marketing data used in this test fails the ANOVA assumption of normality.

```
plot(multi_way_model, 2)
```



Kruskal-Wallis Test

When ANOVA assumptions fail, non-parametric tests such as Kruskal-Wallis can be employed to test for significance.

The Kruskal-Wallis tests below have p-values that are less than the significance level 0.05, this allows the conclusion that there is a significant difference between the treatment groups.

```
kruskal.test(MntMeatProducts ~ NumCatalogPurchases, data=cleaned_df_1)
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  MntMeatProducts by NumCatalogPurchases
## Kruskal-Wallis chi-squared = 1436, df = 11, p-value < 2.2e-16
```

```
kruskal.test(MntMeatProducts ~ Education_1, data=cleaned_df_1)
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  MntMeatProducts by Education_1
## Kruskal-Wallis chi-squared = 7.0293, df = 2, p-value = 0.02976
```

Conclusion

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

The selected marketing variables fail multiple tests for ANOVA assumptions. A better choice for comparison is a non-parametric Kruskal-Wallis test.

The results of the Kruskal-Wallis and Multi-Way-ANOVA both support the hypothesis that the amount spent on meat products is significantly affected by the education and number of catalog purchases made. A Tukey test shows the difference is in the mean amount spent by Bachelors and other groups. In contrast to what was proposed in the hypothesis, there is very little interaction between education and number of catalog purchases, suggesting combinations of these variables do not affect the amount spent on meat.

References

Hult International Business School. (n.d.). marketing data . dataset. retrieved 10/22/22 from https://worldclass.regis.edu/d2l/le/content/297311/Home

MSDS660. (2022). Statistical Methods and Experimental Design. Taught by Dr. Siripun Sanguansintukul.

Statistical Tools for High-Throughput Data Analysis. (N.D.). Two-Way ANOVA Test in R. retrieved 11/19/2022 from sthda.com/english/wiki/two-way-anova-test-in-r

Statistical Tools for High-Throughput Data Analysis. (N.D.). Two-Way ANOVA Test in R. retrieved 11/19/2022 from sthda.com/english/wiki/kruskal-wallis-test-in-r