Week4_Lab

Adam Peetz

MSDS660 Week 4 Assignment

Regis University

Dr. Siripun Sanguansintukul

November 13th 2022

Hypothesis Testing and Confidence Intervals

```
# import libraries
library(tidyverse)
library(data.table)
library(Hmisc)
library(ggpubr)

#heatmap and custom colors
#install.packages("reshape2")
library(reshape2)
#install.packages("viridis")
library("viridis")


#define no_nas function
no_nas = function(x){return(sum(!is.na(x)))}

# import marketing DF
data <- read_csv("marketing.csv",show_col_types = FALSE)

# convert data to table
df<-as.data.table(data)
```

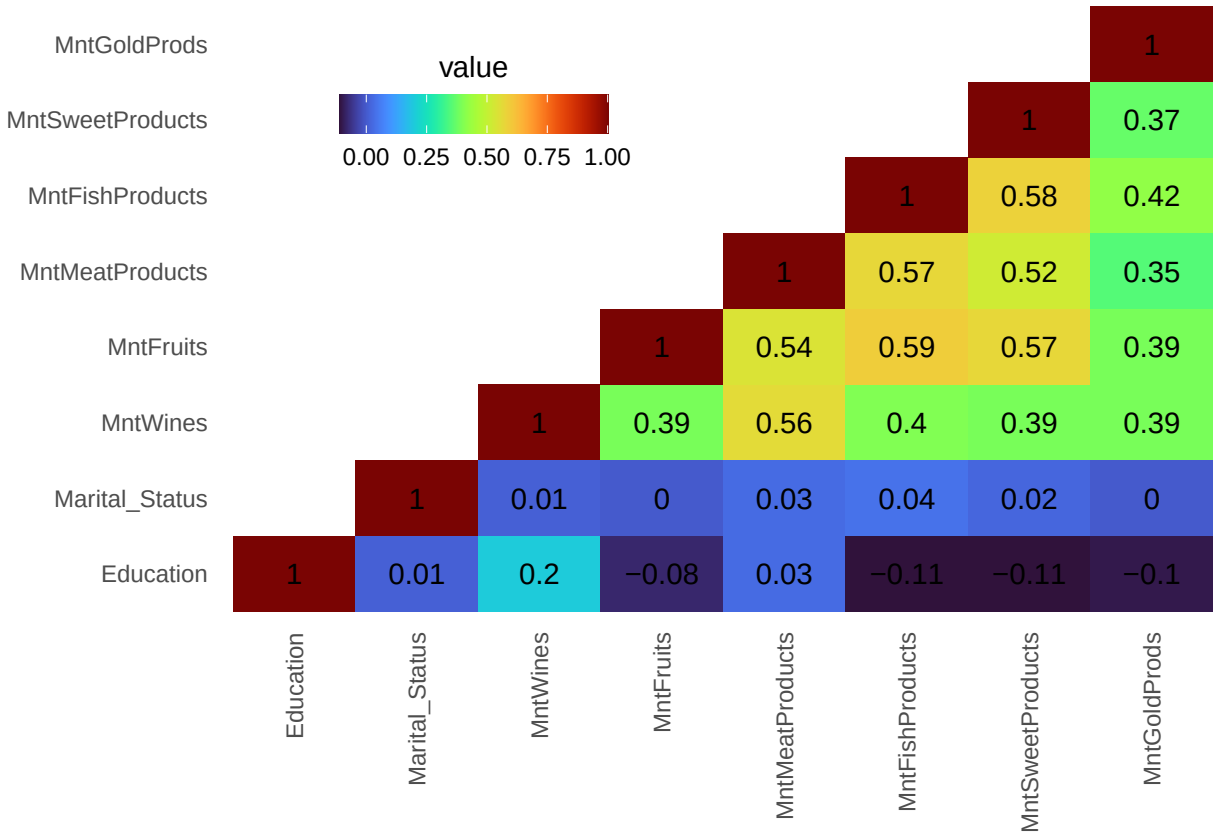
Identifying Correlated Variables: Education and Maritial Status

Marketing data provided by the Hult International School of Business includes demographic information about the education and marital status of customers. There may be correlations between sales amounts and a person's education or marital status. To test this theory, education and marital status are transformed into numerical features and plotted against sales amounts in a correlation heat map. This treatment is shown in the code below and reveals some correlation between a person's education and the amount spent on wine.

```
#create df to hold correlation transformations
corr_df <- df

# convert character to numeric
corr_df$Education <- factor(corr_df$Education)
```

```r
corr_df$Education <- as.numeric(corr_df$Education)

# convert character to numeric
corr_df$Marital_Status <- factor(corr_df$Marital_Status)
corr_df$Marital_Status <- as.numeric(corr_df$Marital_Status)

#subset data to numeric values for sales amounts   and counts
sales_corr_df <- subset(corr_df, select = -c(ID,
                                        Year_Birth,
                                        Income,
                                        Kidhome,
                                        Dt_Customer,
                                        Country,
                                        NumStorePurchases,
                                        Response,
                                        NumCatalogPurchases,
                                        NumWebPurchases,
                                        NumDealsPurchases
                        )               )


#define lower triangle function
get_lower_tri<-function(cormat){
    cormat[lower.tri(cormat)] <- NA
    return(cormat)}

#define upper triangle function
get_upper_tri <- function(cormat){
    cormat[upper.tri(cormat)]<- NA
    return(cormat)}

#translate dataframe to correlation dataframe
cormap <- round(cor(sales_corr_df),2)

#get lower triangle
tri <- get_lower_tri(cormap)

#melt the corrleation dataframe
melted_cormap <- melt(tri, na.rm=TRUE)

#apply gg plotting function
ggplot(data = melted_cormap, aes(x=Var2, y=Var1, fill=value)) +
        geom_tile() +
        geom_text(aes(Var2, Var1, label = value), color = "black", size = 4) +
        scale_fill_viridis(discrete = FALSE, option="H") +
        theme(
        axis.text.x = element_text(angle = 90, vjust = 0.5, hjust=1),
        axis.title.x = element_blank(),
        axis.title.y = element_blank(),
        panel.grid.major = element_blank(),
        panel.border = element_blank(),
        panel.background = element_blank(),
        axis.ticks = element_blank(),
```

```
        legend.justification = c(1, 0),
        legend.position = c(0.4, 0.7),
        legend.direction = "horizontal")+
        guides(fill = guide_colorbar(barwidth = 7, barheight = 1,
                      title.position = "top", title.hjust = 0.5))
```



Corellation: Education and Wine Sales

There may be some correlation between a person's education and the amount spent on wine. This correlation can be proved using t-testing. Before t-testing can be performed the features need to be checked for missing values and treated for outliers.

Outliers and NA's

There are no NA values in the marketing data set. No NA treatment is needed.

Outliers can be checked for with a box and whisker plot. Data points outside the whiskers of the box are considered outlier values. These can be removed using quantile clipping, which removes rows with values above and below the bounds of the whiskers.

```
# check for NA values in education
summary(df$Education)
```

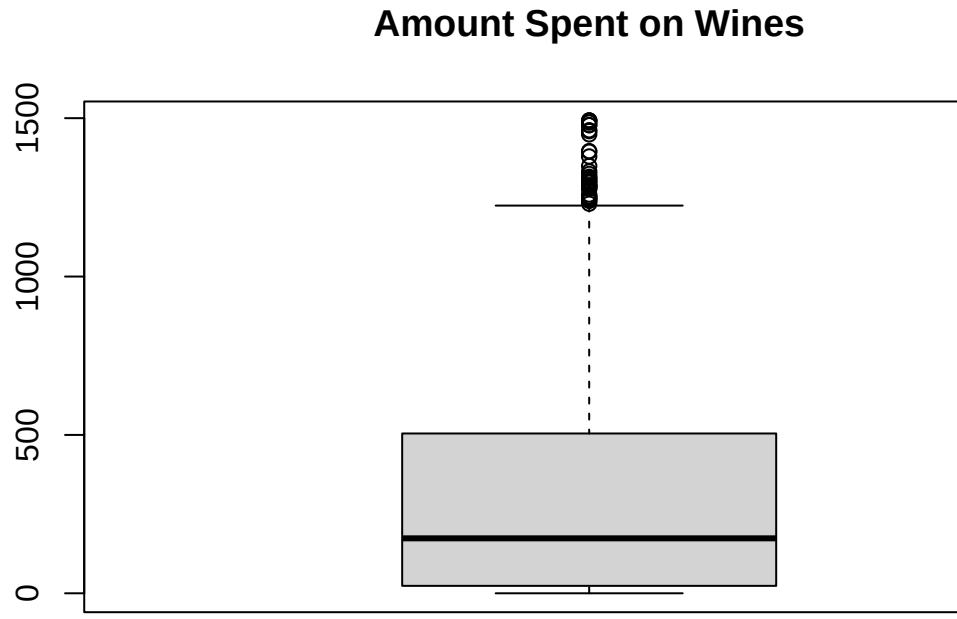
Check for NA Values

```
##    Length     Class      Mode
##      2240 character character
```

```
# check for NA values in amount spent on wine
summary(df$MntWines)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.00   23.75  173.50  303.94  504.25 1493.00
```

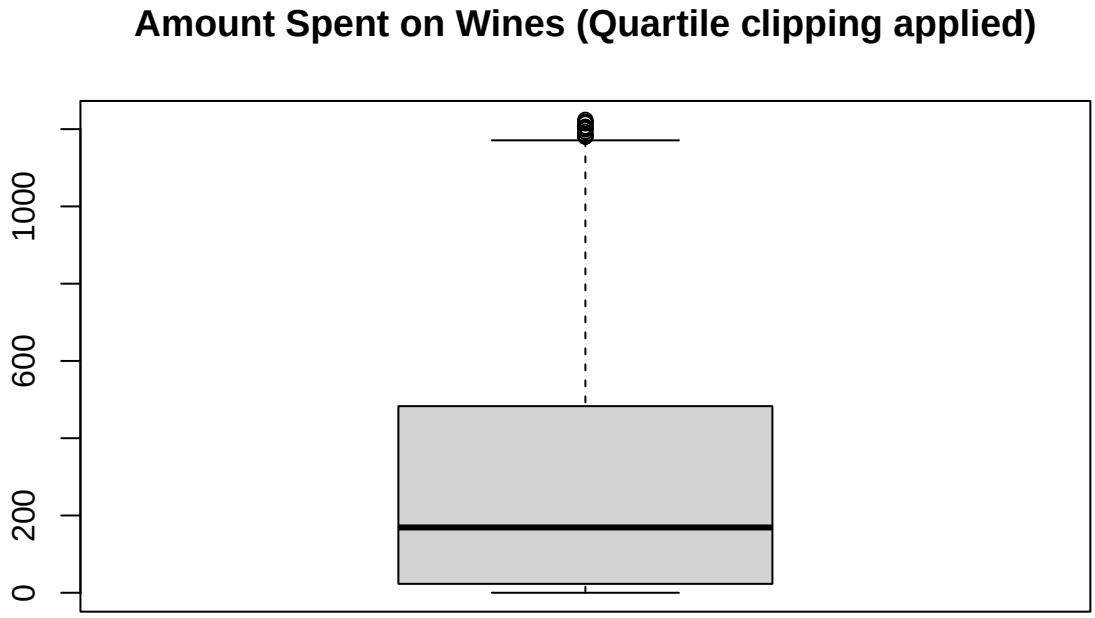
```
# check for outliers
boxplot(df$MntWines, main="Amount Spent on Wines")
```



Check for and remove outliers

```
# remove outliers

# calculate quartiles and inter quartile range
Q <- quantile(df$MntWines, probs=c(.25, .75), na.rm = TRUE)
iqr <- IQR(df$MntWines, na.rm = TRUE)

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

# clip df to remove the rows containing outliers
cleaned_df_1 <- subset(df, df$MntWines > (Q[1] - 1.5*iqr) & df$MntWines < (Q[2]+1.5*iqr))

# visualize the new dataset
```

```
boxplot(cleaned_df_1$MntWines, main="Amount Spent on Wines (Quartile clipping applied)")
```



Compute a CI for a mean difference and a one sample t-test on a numerical variable of interest to you.

The relationship between Education and amount spent on wines will be investigated in more detail. The population will be segmented into groups based on their Education. If education and wine sales are correlated, the population mean of the amount spent on wines by PhD's should be greater than the population mean of Graduate students.

```
#subset to grads
grad_df = cleaned_df_1[cleaned_df_1$Education  %in% c("Graduation"), ]

#subset to phds
phd_df = cleaned_df_1[cleaned_df_1$Education   %in% c("PhD"), ]
```

```
# calculate mean
print("Mean of Graduation Wine Sales")
```

```
## [1] "Mean of Graduation Wine Sales"
```

```
tapply(grad_df$MntWines, grad_df$Education, mean, na.rm=TRUE)
```

```
## Graduation
##   273.7231
```

Confidence Interval

Graduates had a mean spend of 273. A 95% confidence interval suggests the true mean of the population of graduate students is between 247 and 300. If PhD's spend more on wine their population mean should be greater than 300.

```
#calculate confidence interval
xbar = mean(grad_df$MntWines, na.rm=TRUE)
se_xbar = sd(grad_df$MntWines, na.rm=TRUE)/sqrt(no_nas(phd_df$MntWines))
lower = xbar - qt(0.975, df = no_nas(phd_df$MntWines)-1)*se_xbar
upper = xbar + qt(0.975, df = no_nas(phd_df$MntWines)-1)*se_xbar
c(lower, upper)
```

```
## [1] 247.2978 300.1484
```

One Sample T-Test

A one sample one tailed t-test can be used to prove the population mean spend of PhD's is higher than 300 as suggested by the Graduate's confidence interval. It is okay to use a one-tailed test because actually having a population mean higher than 300 would be beneficial to the store.

Hypothesis

Ho: u1 == 300

The population mean of PhD wine sales is equal to 300.

Ha: u1 > 300

The population mean of PhD wine sales greater than 300.

Conclusion

A p-value of < 0.05 allows the rejection of the null hypothesis. The alternate hypothesis is true, the population mean of PhD wine sales is greater than 300.

```
# perform t.test
t.test(phd_df$MntWines, mu=300, alternative ="greater")   ## gives the 95 percent CI as the default
```

```
##
##  One Sample t-test
##
## data:  phd_df$MntWines
## t = 4.1245, df = 466, p-value = 2.2e-05
## alternative hypothesis: true mean is greater than 300
## 95 percent confidence interval:
##  339.871     Inf
## sample estimates:
## mean of x
##  366.4069
```

Two Sample T-Test

The population means of Graduates and PhD's amount wine purchases can be directly checked for similarity using a two sample T-Test.

Hypothesis

Ho: u1 == u2

The population means of Graduate and PhD wine sales is the same.

Ha: u1 != u2

The population means of Graduate and PhD wine sales is not the same.

Conclusion

A p-value of less the 0.05 results in the rejection of the null hypothesis. The alternate hypothesis is true. The population means of Graduate and PhD MntWines is not the same.

```
# perform two sample t.test
t.test(phd_df$MntWines, grad_df$MntWines)
```

```
##
##  Welch Two Sample t-test
##
## data:  phd_df$MntWines and grad_df$MntWines
## t = 5.0646, df = 751.03, p-value = 5.154e-07
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##   56.75791 128.60956
## sample estimates:
## mean of x mean of y
##  366.4069  273.7231
```

Conclusion

There is correlation between a person's education and the amount spent on wine in the marketing data set. A one-sample one-tailed t-test proved the population mean of PhD wine sales was greater than that of Graduate wine sales. A two-sample t-test further proved the population mean of PhD and Graduate wine sales were not the same. The store may want to advertise its selection of wines to PhDs to increase its sales.

References

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

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