MSDS660_Week5_Discussion_Apeetz

2022-11-13

Adam Peetz

MSDS660 Week 4 Discussion

Regis University

Dr. Siripun Sanguansintukul

November 17th 2022

Discussion

Continue working with the loan data set in a different Rmd script.

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.
2. Run a multi-way ANOVA on loan amount received with at least 2 other variables.
3. Is there a significant interaction effect between the levels of each variable? Please plot at least one interaction plot.
4. Test for ANOVA assumptions. (At least the Levene's test for HOV)
5. Does the analysis support the hypothesis you formed initially?
6. Post your rfile and responses to the questions to the Week 5 discussion.

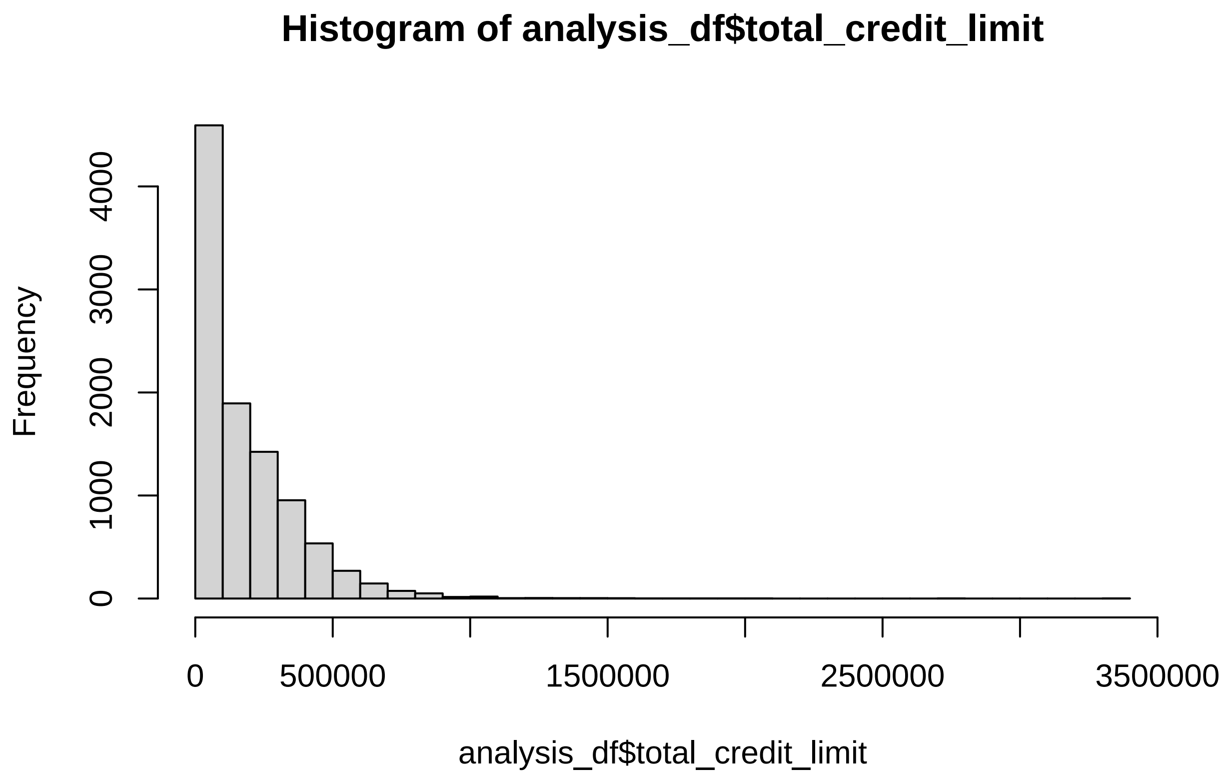
```r
# load libraries
library(tidyverse)
library(data.table)
library(ggpubr)
library(car)
library(dplyr)
library(agricolae)
library(rstatix)

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

Hypothesis

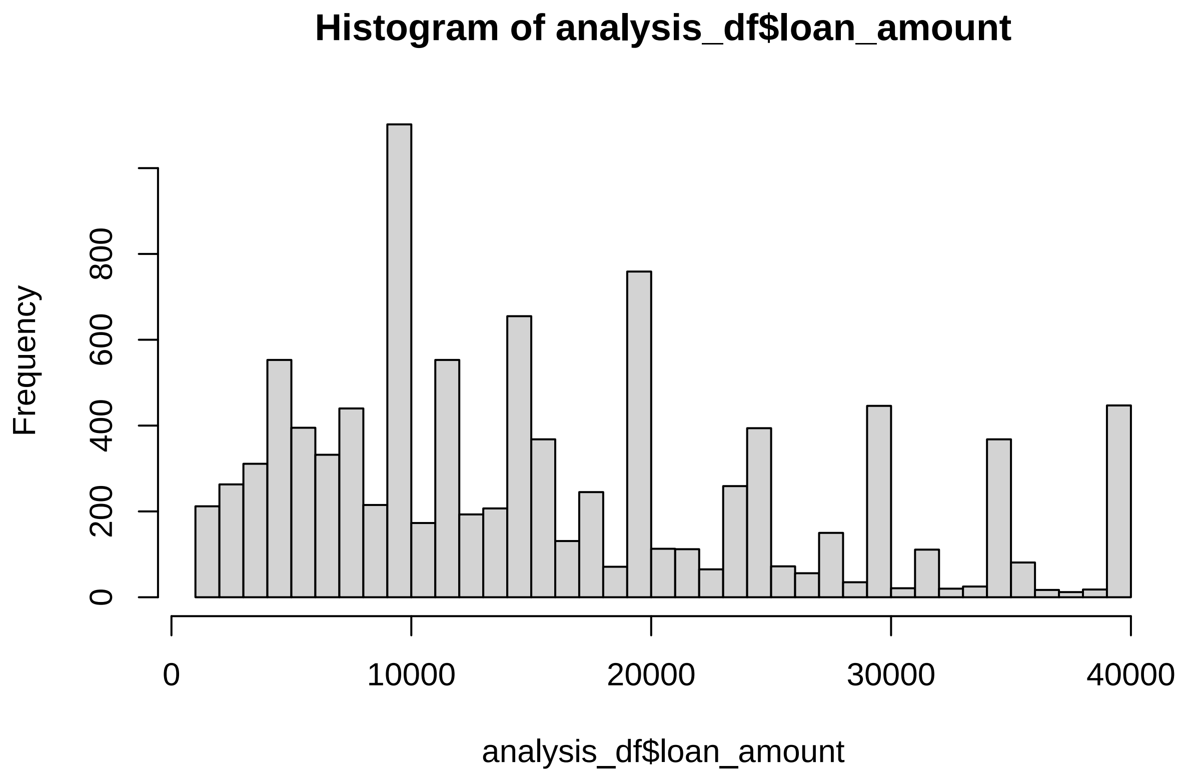
1. Form a hypothesis for the variables that maybe related. You may have both factors and numerical values in your analysis.You would need factors to create an interaction plot.

Hypothesis: Total_credit_limit and application_type will significantly impact loan_amount 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.

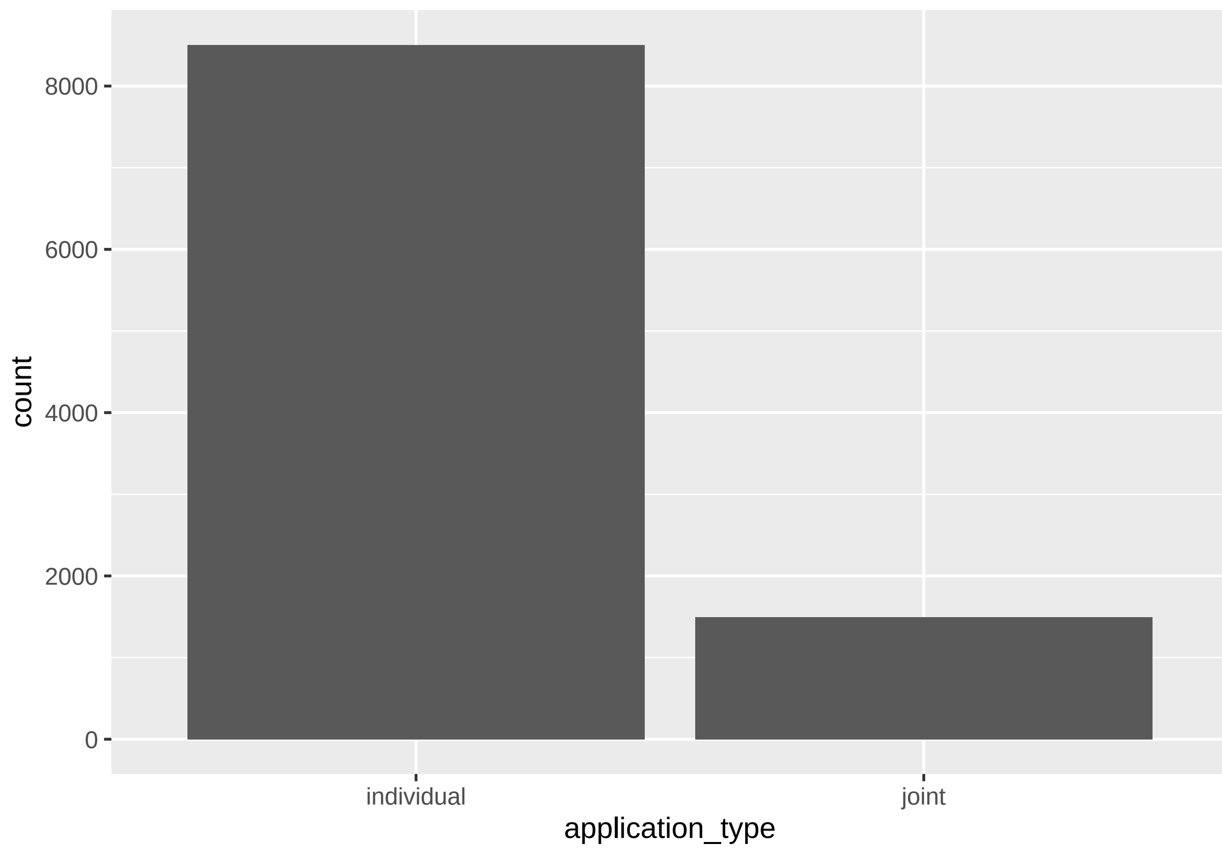
```r
# reduce dataframe to only required variables
analysis_df <- df %>% select(application_type, total_credit_limit, loan_amount)

# eda of selected variables
# histogram
hist(analysis_df$total_credit_limit, breaks=30)
```




```r
# histogram
hist(analysis_df$loan_amount, breaks=30)
```




```r
# count of categorical feature
ggplot(analysis_df, aes(x = application_type)) +
  geom_bar()
```



Levene's test for HOV

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

To perform ANOVA, the data must meet a few assumptions such as homoscedasticity. 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 0.6568 indicates that the null hypothesis is true. All populations variances are equal.

```
# generate levene test
result = leveneTest(loan_amount ~ interaction(application_type, total_credit_limit),
                    data = analysis_df)

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

```
## Levene's Test for Homogeneity of Variance (center = median)
##         Df F value Pr(>F)
## group 9236  0.1646      1
##        763
```

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 total_credit_limit and application_type allow the null hypothesis to be rejected for those variables, there is a significant difference in the mean outcome for these groups. For total_credit_linit and application_type, A p-value of 0.0748 indicates the opposite, the interaction between total_credit_limit and application_type is not significant.

```r
multi_way_model<-aov(loan_amount~total_credit_limit * application_type, data=analysis_df)
summary(multi_way_model)
```

```
##                                       Df    Sum Sq   Mean Sq  F value Pr(>F)
## total_credit_limit                     1 9.759e+10 9.759e+10 1027.154 <2e-16
## application_type                       1 1.357e+10 1.357e+10  142.857 <2e-16
## total_credit_limit:application_type    1 3.017e+08 3.017e+08    3.176 0.0748
## Residuals                           9996 9.497e+11 9.501e+07
##
## total_credit_limit                  ***
## application_type                    ***
## total_credit_limit:application_type .
## Residuals
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

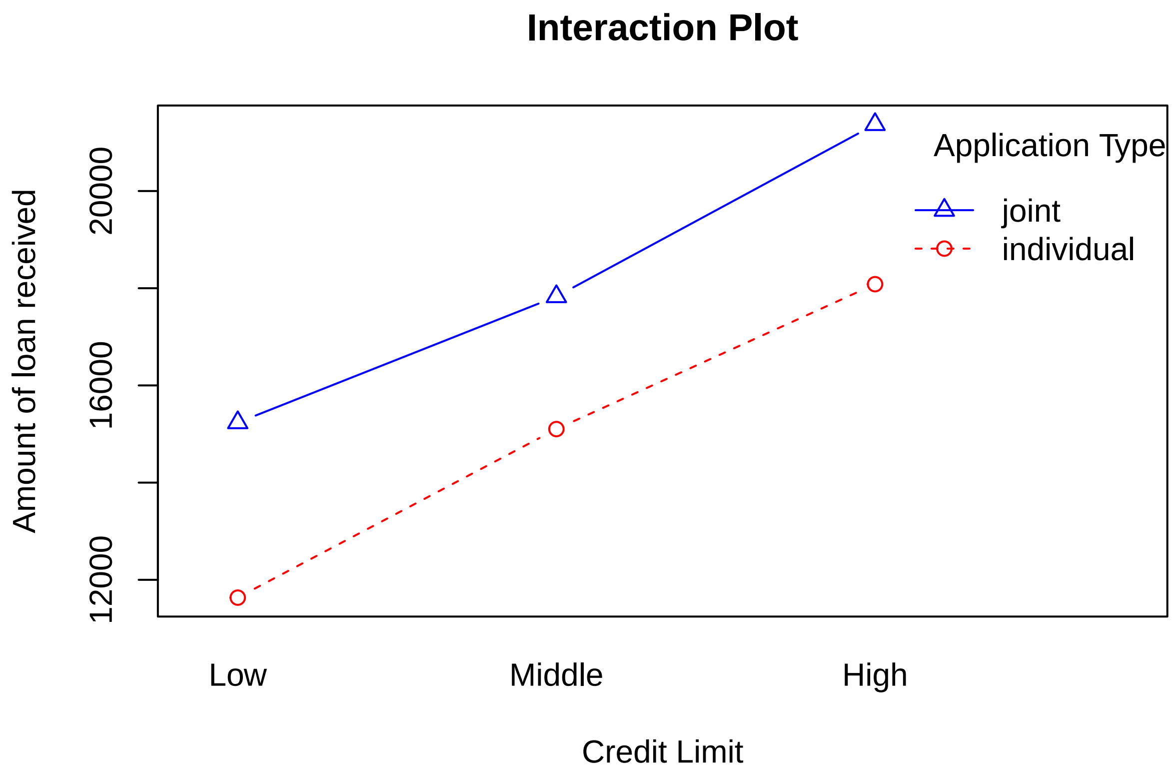
Interaction Plot

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

No significant interaction is shown between total_credit_limit and application_type.

```r
analysis_df <- within(analysis_df, {
  credit_cat <- NA # need to initialize variable
  credit_cat[total_credit_limit < 49999] <- "Low"
  credit_cat[total_credit_limit >= 50000 & total_credit_limit < 99999] <- "Middle"
  credit_cat[total_credit_limit >= 100000] <- "High"
   } )

analysis_df$credit_cat <- factor(analysis_df$credit_cat, levels = c("Low", "Middle", "High"))
```

```r
interaction.plot(x.factor = analysis_df$credit_cat,
                 trace.factor = analysis_df$application_type,
                 response = analysis_df$loan_amount,
                 fun = mean,
                 type = "b",  # shows each point
                 main = "Interaction Plot",
                 legend = TRUE,
                 trace.label = "Application Type",
                 xlab = "Credit Limit",
                 ylab="Amount of loan received",
                 pch=c(1, 2, 3),
                 col = c("Red", "Blue", "Green"))
```



Conclusion

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

The variables loan_amount, total_credit_limit, and application_type are eligible for analysis with anova because they pass the assumption of homoscedasticity as indicated by Levene's test. The results of the multi-way-anova test support the hypothesis that the loan amount is significantly effected by application type and credit limit. However, there was no interaction between application_type and total_credit_limit, suggesting combinations of these variables do not effect the loan amounts given.