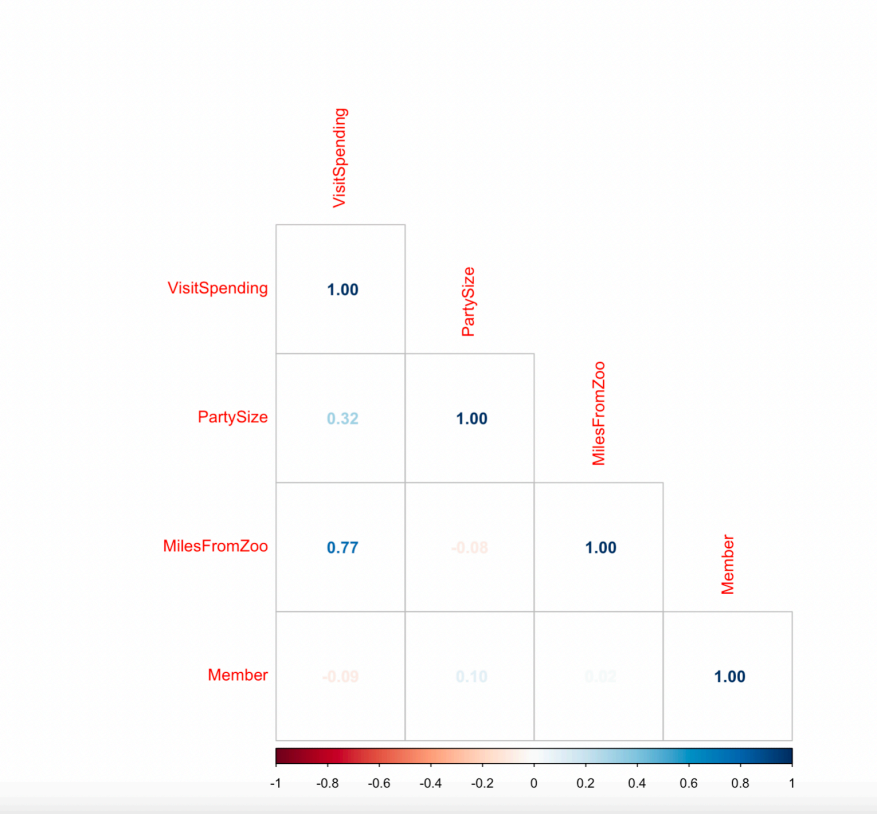
1. From R:
    1. All code from R script

```
# YUEH-TING WU
# MIS 545 Section 02
# Lab05WuY.R
# In this R programming, importing a CSV file and using correlation plot make it
# visualize. And also generate linear regression model to understand the
# coefficients and test for multicollinearity.

# Install the tidyverse, corrplot, and olsrr packages
# install.packages("tidyverse")
# install.packages("corrplot")
# install.packages("olsrr")

# Load the tidyverse, corrplot, and olsrr libraries
library(tidyverse)
library(corrplot)
library(olsrr)

# Set the working directory to my Lab05 folder
setwd("~/MIS 545/Lab05")

# Read ZooVisitSpending.csv into a tibble called zooSpending
zooSpending <- read_csv(file = "ZooVisitSpending.csv",
                        col_names = TRUE,
                        col_types = "niil")

# Display zooSpending in the console
print(zooSpending)

# Display the structure of zooSpending in the console
str(zooSpending)

# Display the summary of zooSpending in the console
summary(zooSpending)

# Recreate the displayAllHistograms() function
displayAllHistograms <- function(tibbleDataset) {
  tibbleDataset %>%
    keep(is.numeric) %>%
    gather() %>%
    ggplot() + geom_histogram(mapping = aes(x=value, fill=key),
                              color = "black") +
    facet_wrap (~key, scales = "free") +
    theme_minimal ()
}
```

```
# Call the displayAllHistograms() function, passing in zooSpending as an
# argument.
displayAllHistograms(zooSpending)

# Display a correlation matrix of zooSpending
cor(zooSpending)

# If the tibble has non-numeric values, limit the correlation matrix to
# numeric values to prevent errors
cor(zooSpending %>% keep(is.numeric))

# And rounded to two decimal places
round(cor(zooSpending), 2)

# Display a correlation plot using the "number" method and limit output to the
# bottom left
corrplot(cor(zooSpending),
         method = "number",
         type = "lower")
# Generate the linear regression model and save it in an object called
# zooSpendingModel
zooSpendingModel <- lm(data = zooSpending,
                       formula = VisitSpending ~ .)

# Display the beta coefficients for the model on the console
print(zooSpendingModel)

# Display the linear regression model results using the summary() function
summary(zooSpendingModel)

# Test for multicollinearity
ols_vif_tol(zooSpendingModel)
```

2. The correlation plot



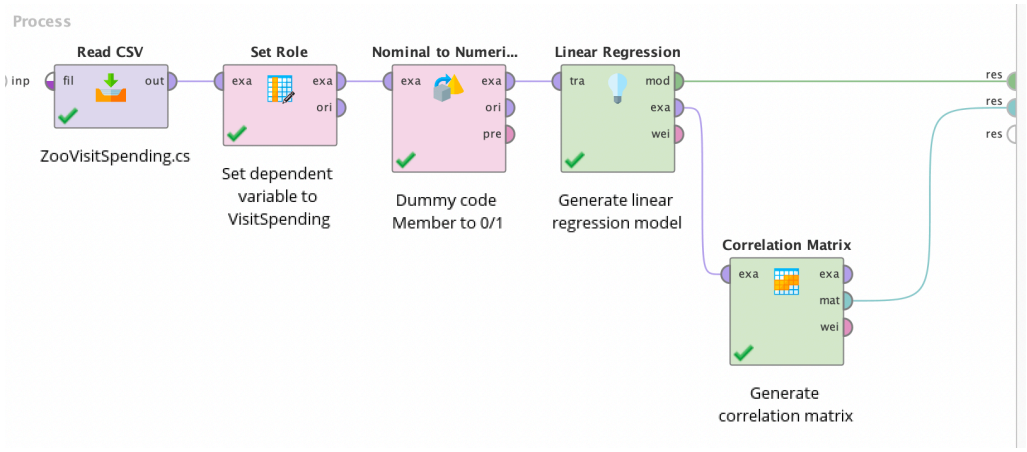
3. The model of summary

```
Call:
lm(formula = VisitSpending ~ ., data = zooSpending)

Residuals:
    Min      1Q  Median      3Q     Max
-57.718 -14.527  -1.476  15.012  54.904

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)    0.22141    6.49061   0.034  0.97284
PartySize      9.13619    1.01756   8.979 4.35e-15 ***
MilesFromZoo   0.88886    0.04865  18.272  < 2e-16 ***
MemberTRUE   -14.90735    4.58300  -3.253  0.00148 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 24.46 on 121 degrees of freedom
Multiple R-squared:  0.765,     Adjusted R-squared:  0.7592
F-statistic: 131.3 on 3 and 121 DF,  p-value: < 2.2e-16
```

4. The result from the test for multicollinearity

```
> # Test for multicollinearity
> ols_vif_tol(zooSpendingModel)
     Variables Tolerance      VIF
1    PartySize 0.9831086 1.017182
2 MilesFromZoo 0.9926983 1.007355
3   MemberTRUE 0.9890274 1.011094
```

2. From RapidMiner:
    1. A screenshot of your process



    2. The correlation matrix

Attribu...	Membe...	PartySi...	MilesFr...	VisitSp...
Membe...	1	0.101	0.021	-0.087
PartySize	0.101	1	-0.080	0.320
MilesFro...	0.021	-0.080	1	0.773
VisitSpe...	-0.087	0.320	0.773	1

3. A screenshot of the linear regression model results

Attribute	Coefficient	Std. Error	Std. Coefficient	Tolerance	t-Stat	p-Value	Code
Member = 1	-14.907	4.583	-0.144	0.996	-3.253	0.001	***
PartySize	9.136	1.018	0.399	0.991	8.979	0.000	****
MilesFromZoo	0.889	0.049	0.808	0.993	18.272	0	****
(Intercept)	0.221	6.491	?	?	0.034	0.973	

3. Answer the following question in a sentence: Within the model, which variables are statistically significant?

In the linear regression model, the PartySize, the MilesFromZoo, and the Member are statistically significant.

4. Answer the following question in a sentence: How much of the variance in zoo spending can be explained by the variance in party size, miles from the zoo, and zoo membership?

There are 76.5% of the variance in zoo spending can be explained by all the independent variables.

5. Answer the following question in a sentence: Within the model, how much more/ less will zoo spending be with each additional guest in a party?

In this model, for each additional guest in a party, we can expect that the zoo spending will be increased by 9.13619.

6. Answer the following question in a sentence: Within the model, how much more/ less is zoo spending for members compared with non-members? Explain why this might be the case.

In this model, comparing with non-members, we can expect that the zoo spending for members will be decreased by 14.90735. I predict that those members might very like to go to the zoo. So, they might be already paid an annual pass for the zoo. Thus, the average price of the zoo pass might be less than non-members by 14.90735.

7. Answer the following question in a sentence: Within the model, how much more/ less will spending be for each additional mile travelled to visit the zoo? Explain why this might be the case.

In this model, for each additional mile travelled to visit the zoo, we can expect that the zoo spending will be increased by 0.88886. I predict that the more miles they far from the zoo, then the more fuel or the more transportation fees they need to use. So, the increase of money might be the consumption of fuel or transportation fees.

8. Answer the following question in a sentence: Does the model suffer from multicollinearity? If so, what could be done to rectify it? If not, why?

There is no multicollinearity in this model because none of the tolerance values is less than 0.2 or VIF is greater than 5.