MIS 545 Lab 05 Data Preprocessing

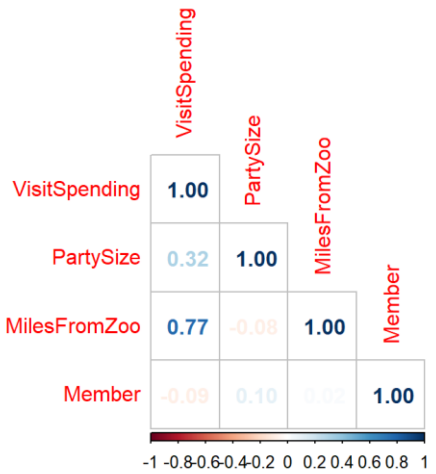
1.R Script-

```
# LIKITHA TRIPURANENI, VIHARI REDDY TUMMURU

# MIS 545 01

# Lab05Group01TripuraneniTummuru.R

# This code demonstrates the summary of ZooVisitSpending.csv data, displays all

# histograms, correlation matrix and plot, generates linear regression model

# and tests for multicollinearity


# Install tidyverse, corrplot, olsrr packages -----------------------------

# install.packages("tidyverse")

# install.packages("corrplot")

# install.packages("olsrr")


# Load tidyverse, corrplot, olsrr packages --------------------------------

library("tidyverse")

library("corrplot")

library("olsrr")


# Set the current working directory ---------------------------------------

setwd("C:/MIS54501LAB05")


# Read the ZooVisitSpending.csv file into a tibble ------------------------

zooSpending <- read_csv(file= "ZooVisitSpending.csv",

                       col_types = "niil",

                       col_names = TRUE)


# Display zooSpending tibble -----------------------------------------------

print(zooSpending)


# Display the structure of zooSpending -------------------------------------
```

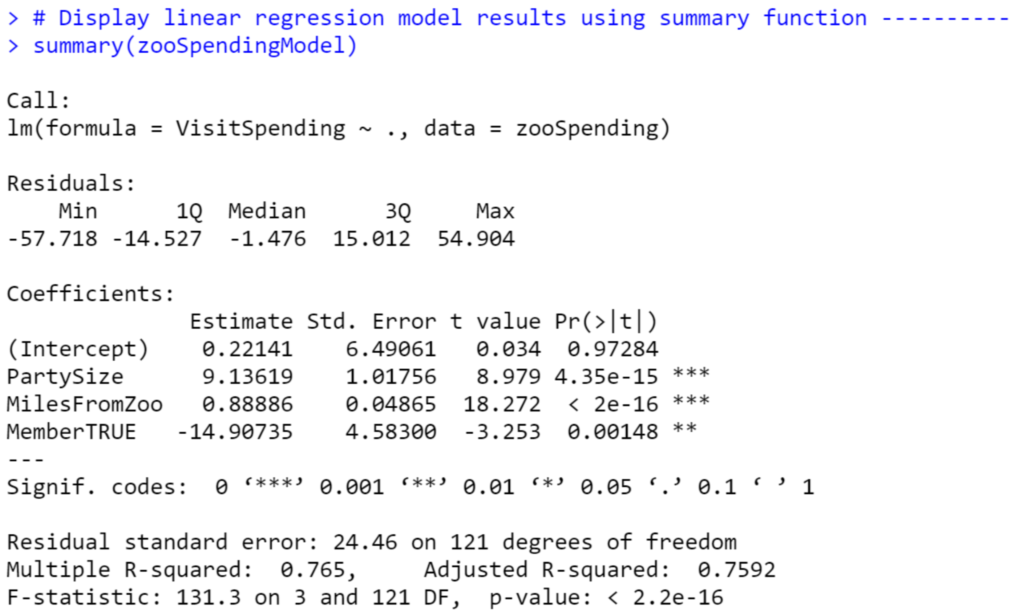
```
print(str(zooSpending))

# Summarize the zooSpending tibble -----------------------------------------
print(summary(zooSpending))

# Create displayAllHistograms function ------------------------------------
displayAllHistograms <- function(tibbleDataset) {
  tibbleDataset %>%
    keep(is.numeric) %>%
    gather() %>%
    ggplot() + geom_histogram(mapping = aes(x=value,fill=key),
                              color="black",
                              #bins=30
                              )+
    facet_wrap(~key,scales="free")+
    theme_minimal()
}

# Display all histograms for zooSpending ----------------------------------
displayAllHistograms(zooSpending)

# Display Correlation Matrix ----------------------------------------------
cor(zooSpending)

# Limit the correlatin matrix to numeric values to prevent errors ---------
cor(zooSpending %>% keep(is.numeric))

# Round the correlation matrix to two decimals ----------------------------
round(cor(zooSpending),2)

# Display correlation plot ------------------------------------------------
```

```
corrplot(cor(zooSpending),
         method = "number")

# Limit correlation plot to the bottom left -------------------------------
corrplot(cor(zooSpending),
         method = "number",
         type = "lower")

# Generate linear regression model ----------------------------------------
zooSpendingModel <- lm(data = zooSpending,
                       formula = VisitSpending ~ .)

# Display the beta coefficients -------------------------------------------
print(zooSpendingModel)

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

# Test for multicollinearity ----------------------------------------------
ols_vif_tol(zooSpendingModel)
```

Correlation plot:

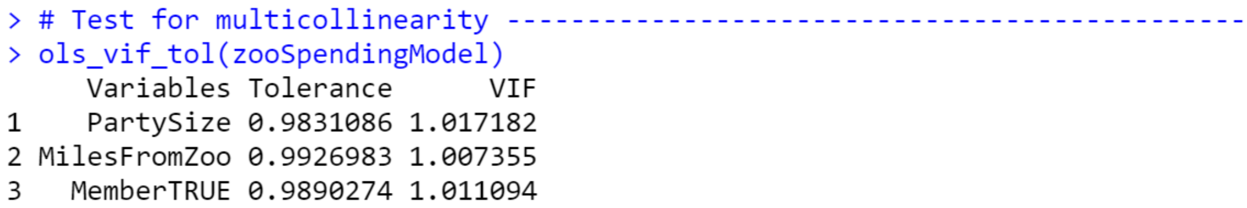


Model Summary:



```
> # Display linear regression model results using summary function ----------
> summary(zooSpendingModel)

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
```

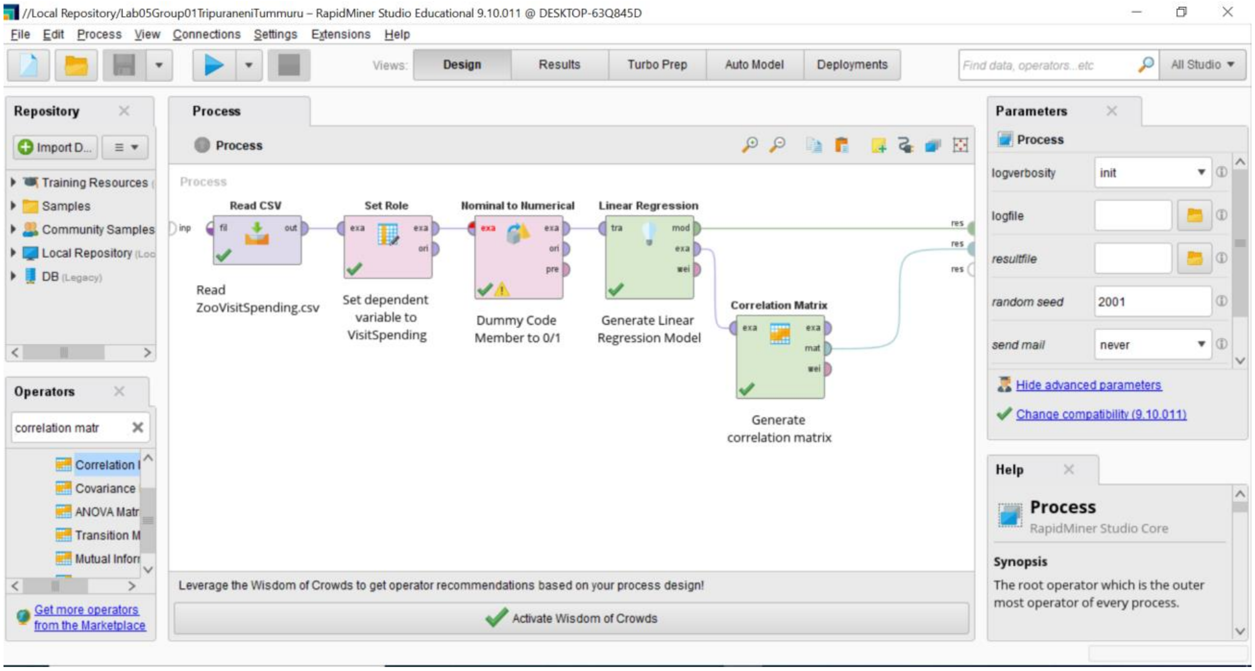
Test for collinearity:



```
> # 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.Rapid Miner-

Process:



//Local Repository/Lab05Group01TripuraneniTummuru – RapidMiner Studio Educational 9.10.011 @ DESKTOP-63Q845D

File Edit Process View Connections Settings Extensions Help
Views: Design Results Turbo Prep Auto Model Deployments
Find data, operators...etc
All Studio
Repository
Import D...
Training Resources
Samples
Community Samples
Local Repository
DB (Legacy)
Operators
correlation matr
Correlation
Covariance
ANOVA Matr
Transition M
Mutual Infor
Get more operators from the Marketplace
Process
Read CSV
Read ZooVisitSpending.csv
Set Role
Set dependent variable to VisitSpending
Nominal to Numerical
Dummy Code Member to 0/1
Linear Regression
Generate Linear Regression Model
Correlation Matrix
Generate correlation matrix
Parameters
Hide advanced parameters
Change compatibility (9.10.011)
Help
Process
RapidMiner Studio Core
Activate Wisdom of Crowds

logverbosity: init

logfile:

resultfile:

random seed: 2001

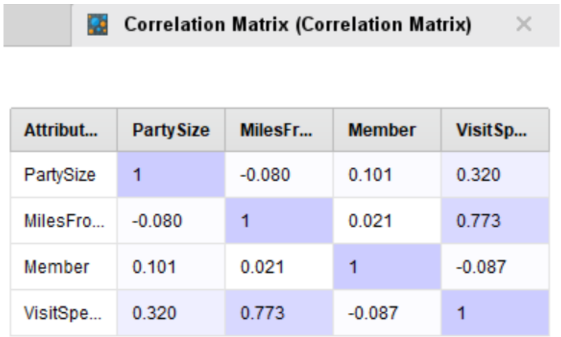
send mail: never

Synopsis

The root operator which is the outer most operator of every process.

Leverage the Wisdom of Crowds to get operator recommendations based on your process design!

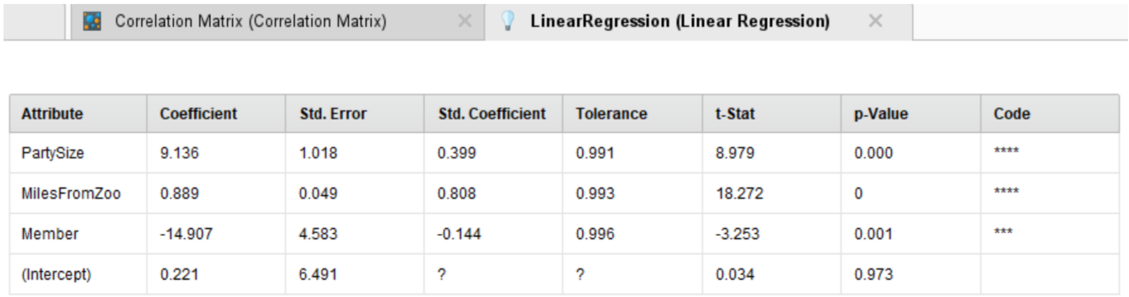
Results: Correlation Matrix



Correlation Matrix (Correlation Matrix)

Attribut...	PartySize	MilesFr...	Member	VisitSp...
PartySize	1	-0.080	0.101	0.320
MilesFro...	-0.080	1	0.021	0.773
Member	0.101	0.021	1	-0.087
VisitSpe...	0.320	0.773	-0.087	1

Results: Linear Regression Model



Correlation Matrix (Correlation Matrix) | LinearRegression (Linear Regression)

Attribute	Coefficient	Std. Error	Std. Coefficient	Tolerance	t-Stat	p-Value	Code
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	****
Member	-14.907	4.583	-0.144	0.996	-3.253	0.001	***
(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?

Answer: MilesFromZoo, PartySize and Memeber 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?

Answer: 75.92% of the varaince in zoo spending can be explained.

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?

Answer: Zoo spending will be 9.136 more for each additional guest.

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.

Answer: Zoo spending will be 14.907 less for members compared to non-members. This is because in the linear regression model shows the coefficient for Member to be -14.907.

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.

Answer: Zoo spending will be 0.889 more for each additional mile travelled. This is because in the linear regression model shows the coefficient for MilesFromZoo to be 0.889.

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?

Answer: The model doesn’t suffer from multicollinearity because for all the values- VIF < 5.0 and Tolerance > 0.2