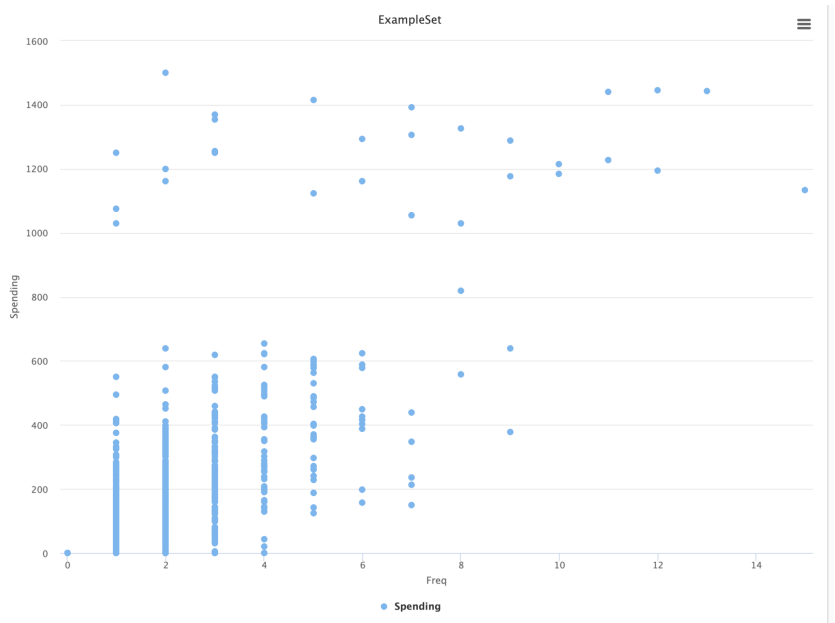
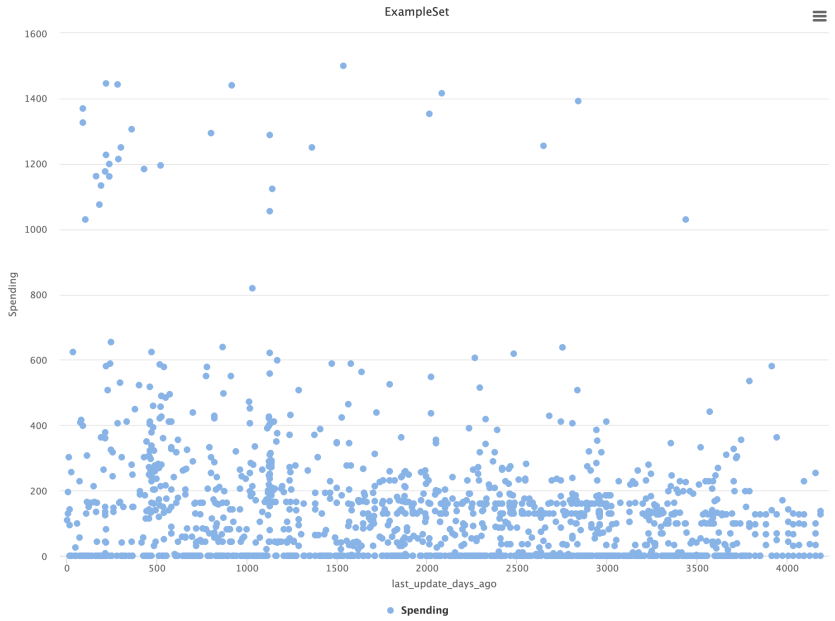
Exploration

a) Explore the relationship between Spending and each of the two continuous variables by creating two scatters plots (SPENDING vs. FREQ and SPENDING vs. LAST_UPDATE). Does there seem to be a linear relationship there? => **Capture Screen** !



มีความเป็น linear relationship



ไม่มีความเป็น linear relationship

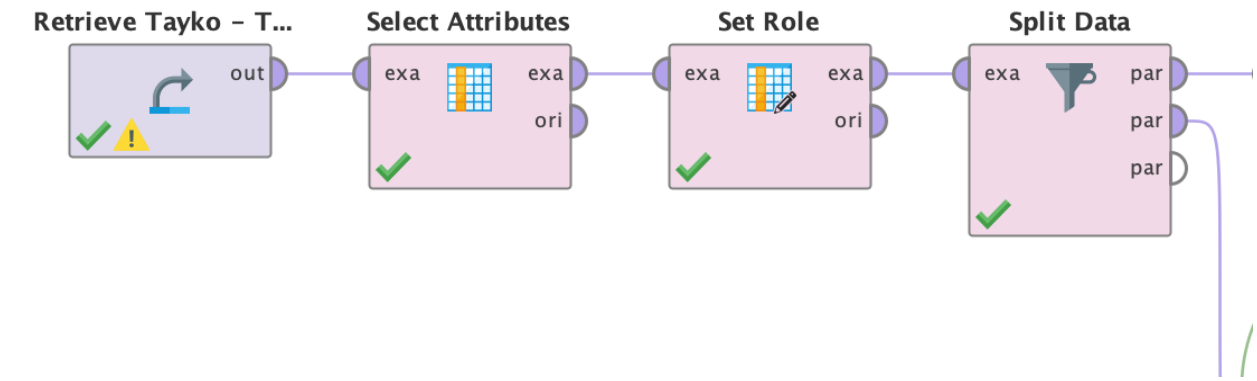
หา correlation matrix มายืนยันผลลัพธ์ว่าสิ่งที่ทำนายเป็นจริงไหม

Attribu...	Freq	last_up...	Spendi...
Freq	1	-0.348	0.691
last_up...	-0.348	1	-0.257
Spending	0.691	-0.257	1

สังเกตได้ว่า Freq กับ Spending มี positive relationship กันจริง

b) Fit a predictive model for SPENDING using only the following predictors: Freq, Last_update, Web_order, Gender, US, Adress_is_res [Use all these features]

1) Partition the 1000 records into training (Partition=t) & test sets (Partition=v)



จาก dataset ทั้งหมดทำการแบ่ง dataset เป็นสองส่วน ส่วนที่หนึ่งคือ 60% เป็นข้อมูลของการ train และ 40% เป็นข้อมูลของการ test

2) Run a multiple regression model for SPENDING with the 6 predictors. => Give the regression equation 1

Attribute	Coefficient
US	-4.342
Freq	96.863
last_update_days_ago	-0.008
Web order	17.206
Address_is_res	-93.120
(Intercept)	0.732

จะได้สมการของ regression1 คือ

Spending = -4.342*US + 96.863*Freq -0.008*last_update + 17.206*Web order - 93.120*Address_is_res + 0.732

LinearRegression

```
- 4.342 * US
+ 96.863 * Freq
- 0.008 * last_update_days_ago
+ 17.206 * Web order
- 93.120 * Address_is_res
+ 0.732
```

3) Based on the above regression equation and P-value of each predictor, identify the characteristics of high spending buyers.? Please justify your answer

Attribute	Coefficient	Std. Error	Std. Coefficient	Tolerance	t-Stat	p-Value
US	-4.342	10.508	-0.009	1.000	-0.413	0.680
Freq	96.863	3.158	0.689	0.926	30.676	0
last_update_days_ago	-0.008	0.004	-0.046	0.919	-2.070	0.039
Web order	17.206	8.187	0.044	0.990	2.102	0.036
Address_is_res	-93.120	10.058	-0.200	0.957	-9.258	0
(Intercept)	0.732	14.415	?	?	0.051	0.960

คนที่มีการซื้อสูงจะมีลักษณะก็คือ จะมีการซื้อที่ถี่ รวมทั้ง last_update กับ web order จะสำคัญเพราะซื้อบ่อยทำให้สองค่านี้ส่งผล และที่สำคัญคนที่ซื้อเยอะๆส่วนใหญ่จะไม่ได้ซื้อเข้าบ้านตัวเอง

4) If we need to reduce the number of predictors, which predictor(s) would be dropped from the model?

จะดรอป attribute ของ US กับ gender เพราะมีว่าค่า p value ที่สูง

Fitting second model

c) Fit a second predictive model for SPENDING using your best predictors:

1) Apply multiple linear regression to create a spending prediction model. Then, give the regression equation 2.

Attribute	Coefficient	Std. Error	Std. Coefficient	Tolerance	t-Stat	p-Value
Freq	96.818	3.155	0.689	0.927	30.691	0
last_update_days_ago	-0.008	0.004	-0.047	0.919	-2.103	0.036
Web order	17.236	8.184	0.044	0.990	2.106	0.035
Address_is_res	-93.278	10.048	-0.200	0.957	-9.284	0
(Intercept)	-2.512	12.085	?	?	-0.208	0.835

ได้สมการดังนี้

> **LinearRegression**
>
> ```
>   96.818 * Freq
> - 0.008 * last_update_days_ago
> + 17.236 * Web order
> - 93.278 * Address_is_res
> - 2.512
> ```

2) Displays the prediction results of the purchase amount in the first record of the test data set, along with indicating the error obtained.

Row No.	Spending	prediction(...	Freq	last_updat...	Web order	Gender=m...	Address_is...
1	0	-8.905	0	2900	1	1	0

Error ที่เกิดขึ้น -8.905

3) Give the performance of the model (error) on the test data set.