Predicting software reselling profit

Tayko Software is a software catalog firm that sells games and educational software. It has recently put together a revised collection of items in a new catalog, which it mailed out to its customers. This mailing yielded 1000 purchases. Based on these data, Tayko wants to devise a model for predicting the spending amount that a purchasing customer will yield.
The file Tayko.xls contains the following attributes:

1.	US	Is it a US address?	binary	1: yes 0: no
2 - 16	Source_*	Source catalog for the record (15 possible sources)	binary	1: yes 0: no
17.	Freq.	Number of transactions in last year at source catalog	numeric	
18.	last_update_days_ago	How many days ago was last update to cust. record	numeric	
19.	1st_update_days_ago	How many days ago was 1st update to cust. record	numeric	
20.	Web_order	Customer placed at least 1 order via web	binary	1: yes 0: no
21.	Gender=mal	Customer is male	binary	1: yes 0: no
22.	Address_is_res	Address is a residence	binary	1: yes 0: no
23.	Purchase	Person made purchase in test mailing	binary	1: yes 0: no
24.	Spending	Amount spent by customer in test mailing ($)	numeric	
25.	Partition	Variable indicating which partition the record will be assigned to	alpha	t: training v: validation

In this study, we are interested only on the purchases (**Purchase=1**). All dummy variables are already created !

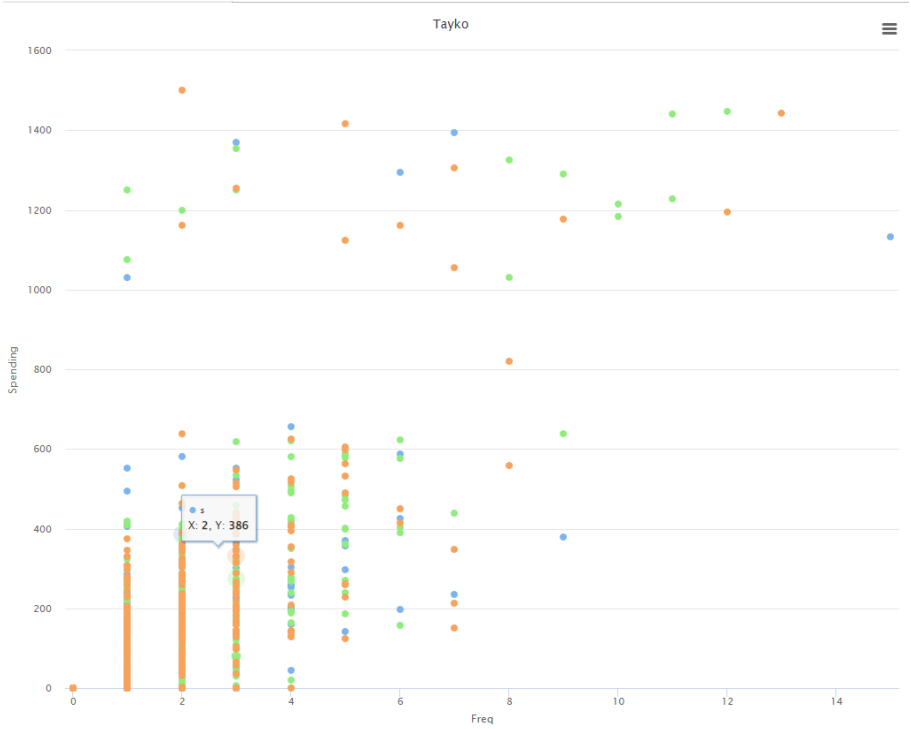
Exploration

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** !

SPENDING vs. FREQ

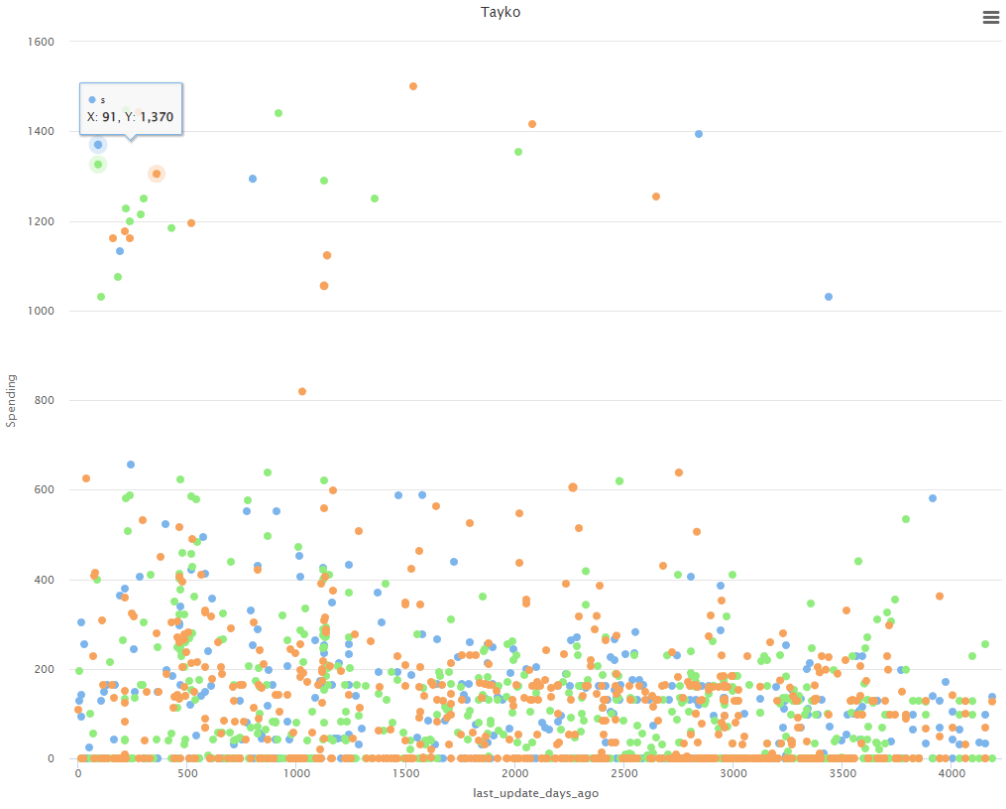
seem to be a linear relationship in this relationship.

(direct variation)



SPENDING vs. LAST_UPDATE

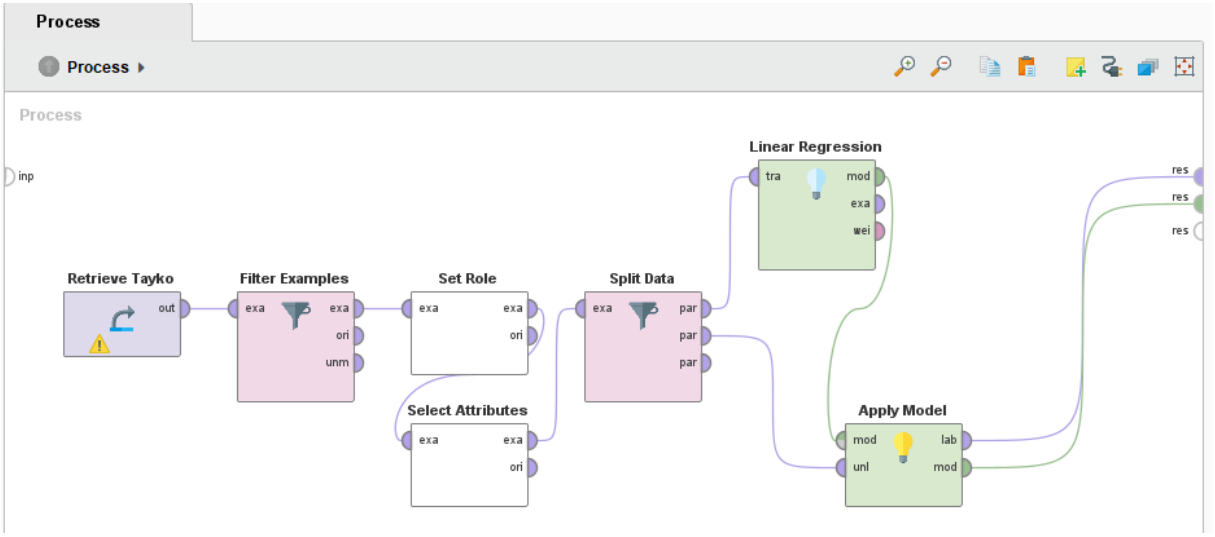
Not linear relationship



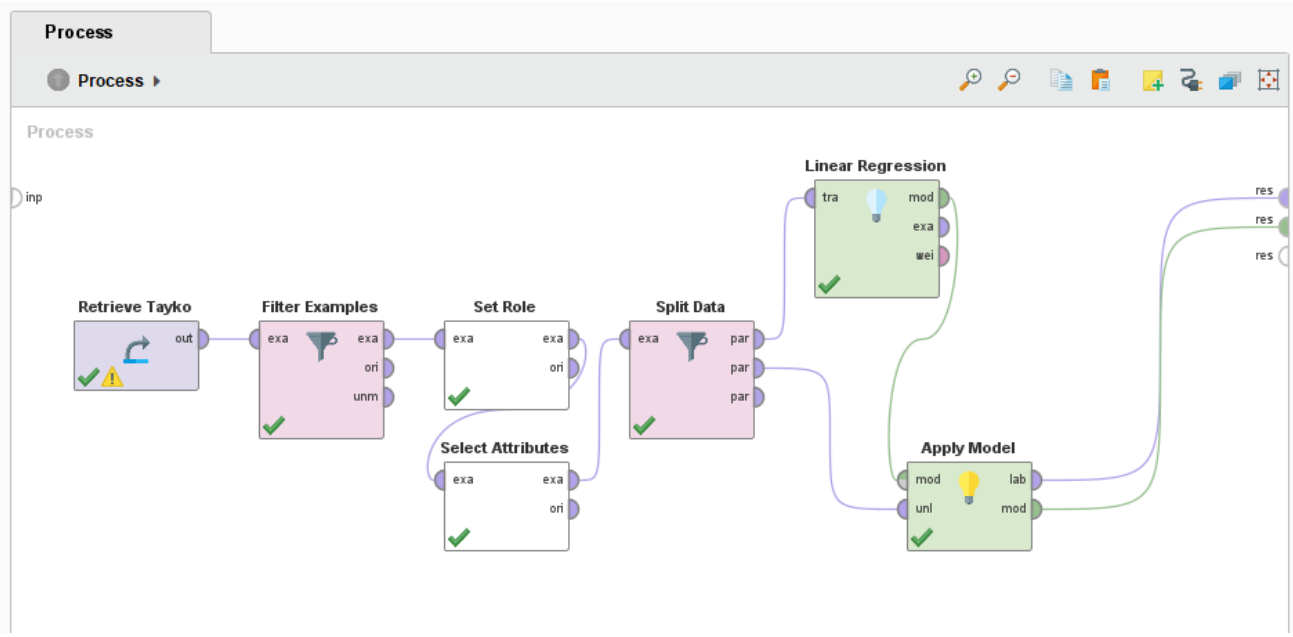
Fitting first model

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)



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



Y= - 11.867 * US
+ 91.373 * Freq
- 0.017 * last_update_days_ago
+ 5.045 * Web order
- 1.670 * Gender=male
- 101.621 * Address_is_res
+ 83.624

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

Address is not resident and has low number of transaction in last year at source catalog is high spending buyers.

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

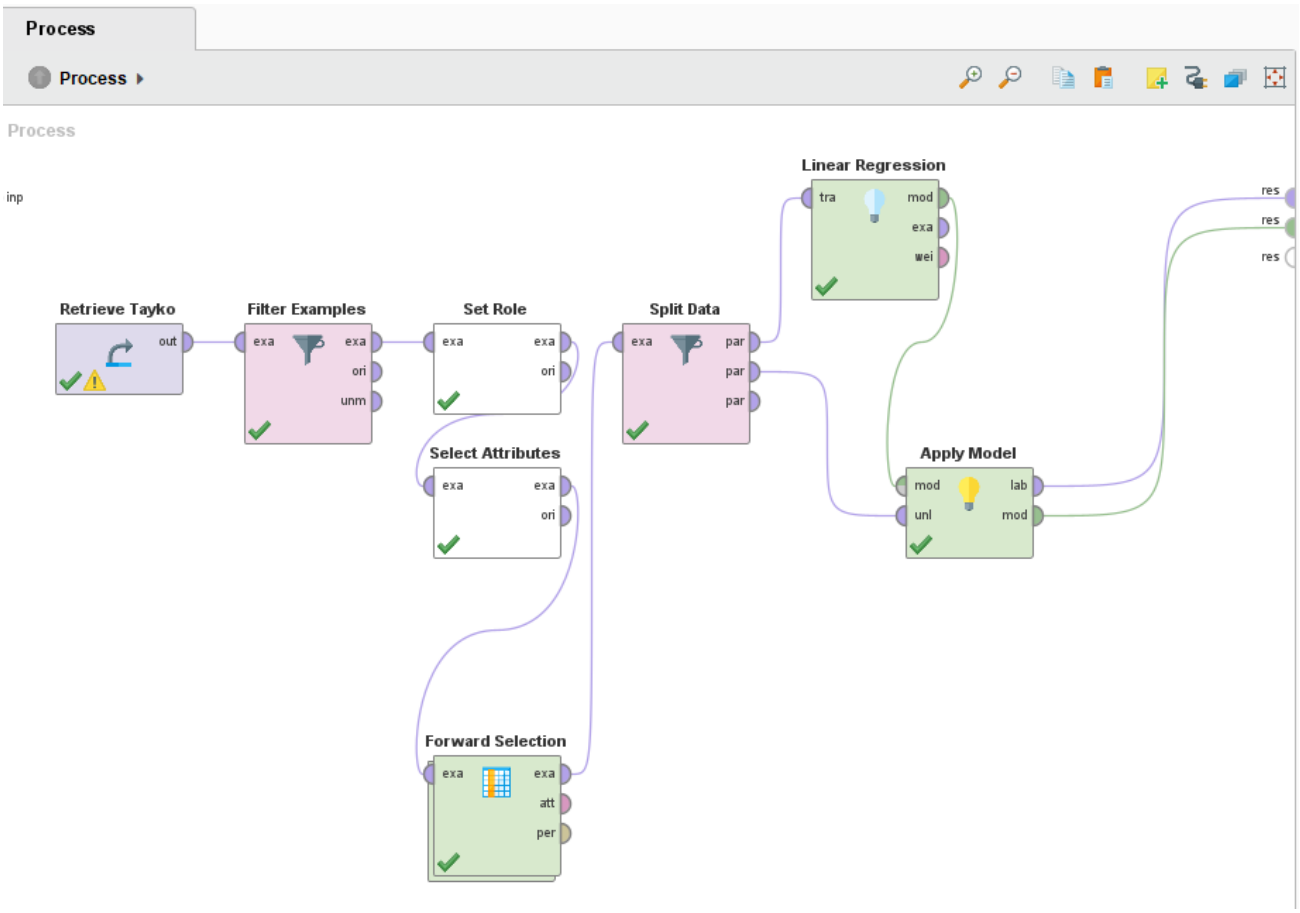
Attribute	Coefficient	Std. Error	Std. Coefficient	Tolerance	t-Stat	p-Value	Code
US	-11.867	16.120	-0.019	0.999	-0.736	0.462	
Freq	91.373	3.932	0.655	0.921	23.237	0	****
last_update_days_ago	-0.017	0.006	-0.084	0.901	-3.004	0.003	***
Web order	5.045	12.092	0.011	1.000	0.417	0.677	
Gender=male	-1.670	12.134	-0.004	1.000	-0.138	0.891	
Address_is_res	-101.621	14.698	-0.185	0.967	-6.914	0.000	****
(Intercept)	83.624	21.996	?	?	3.802	0.000	****

Us, Gender and WebOrder because the p-Value is outstanding from the others

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.



- 27.441 * US
+ 85.356 * Freq
+ 6.832 * Web order
+ 45.486

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(S...	US	Freq	Web order
1	489	366.299	1	4	1

Percentage error = [(489 - 366.299) /366.299]*100 = 33.49%

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

root_mean_squared_error

```
root_mean_squared_error: 175.434 +/- 0.000
```