Question 1:

For every day, if I want to know how Covid-19 impacts us, or if I want to make some investment, we need some data to know these. So, here is Covid-19 updated information and stock price updated information.

Data source for Covid-19 data:

	Province/State	Country/Region	Lat	Long	1/22/20	1/23/20	1/24/20	1/25/20	1/26/20	1/27/20	...	8/30/20	8/31/20	9/1/20	9/2/20	9/3/20	9/4/20	9/5/20	9/6/20	9/7/20	9/8/20
0	NaN	Afghanistan	33.93911	67.709953	0	0	0	0	0	0	...	38162	38165	38196	38243	38288	38304	38324	38398	38494	38520
1	NaN	Albania	41.15330	20.168300	0	0	0	0	0	0	...	9380	9513	9606	9728	9844	9967	10102	10255	10406	10553
2	NaN	Algeria	28.03390	1.659600	0	0	0	0	0	0	...	44146	44494	44833	45158	45469	45773	46071	46364	46653	46938
3	NaN	Andorra	42.50630	1.521800	0	0	0	0	0	0	...	1124	1176	1184	1199	1199	1215	1215	1215	1261	1261
4	NaN	Angola	-11.20270	17.873900	0	0	0	0	0	0	...	2624	2654	2729	2777	2805	2876	2935	2965	2981	3033

5 rows × 235 columns

The following 3 datasets includes province/states, country/region, latitude and longitude of the region, and daily reports.

https://raw.githubusercontent.com/CSSEGISandData/COVID-19/master/csse_covid_19_data/csse_covid_19_time_series/time_series_covid19_confirmed_global.csv

In each day, this dataset will be updated with the total number of confirmed people in the world during Covid-19.

https://raw.githubusercontent.com/CSSEGISandData/COVID-19/master/csse_covid_19_data/csse_covid_19_time_series/time_series_covid19_deaths_global.csv

In each day, this dataset will be updated with the total number of deaths in the world during Covid-19.

https://raw.githubusercontent.com/CSSEGISandData/COVID-19/master/csse_covid_19_data/csse_covid_19_time_series/time_series_covid19_recovered_global.csv

In each day, this dataset will be updated with the total number of recovered people in the world during Covid-19.

Data source for Yahoo stock price:

Load the data

```
import pandas_datareader as web
df = web.DataReader('AAPL', data_source = 'yahoo', start = '2010-01-01', end = '2020-09-09')
#df.head(5)
df.tail(12)
#df.shape
```

Out[5]:

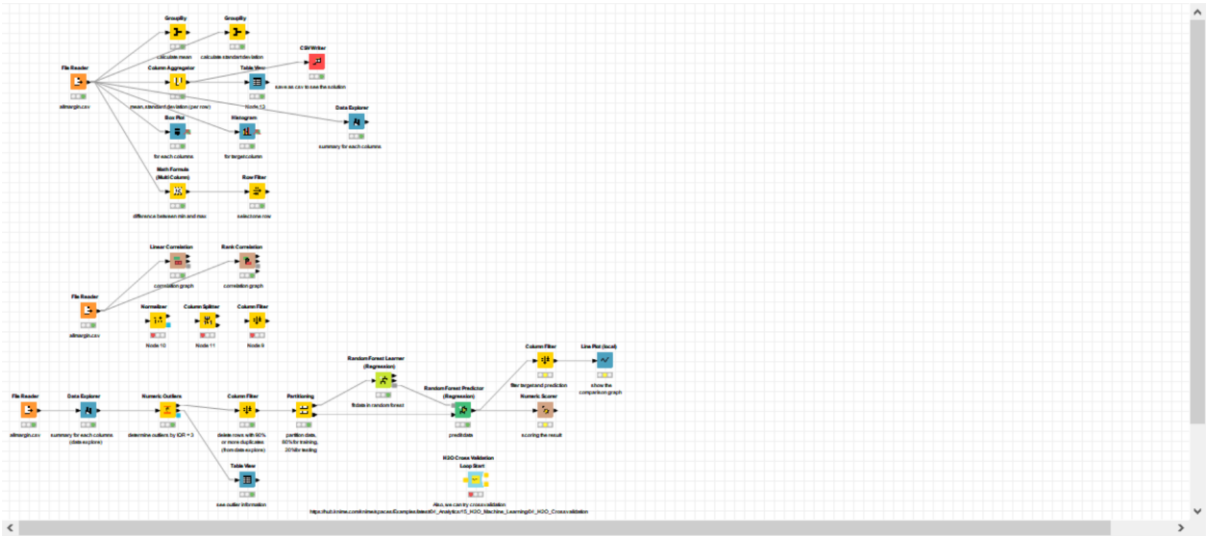
Date	High	Low	Open	Close	Volume	Adj Close
2020-08-24	128.785004	123.937500	128.697495	125.857498	345937600.0	125.857498
2020-08-25	125.180000	123.052498	124.697502	124.824997	211495600.0	124.824997
2020-08-26	126.992500	125.082497	126.180000	126.522499	163022400.0	126.522499
2020-08-27	127.485001	123.832497	127.142502	125.010002	155552400.0	125.010002
2020-08-28	126.442497	124.577499	126.012497	124.807503	187630000.0	124.807503
2020-08-31	131.000000	126.000000	127.580002	129.039993	225702700.0	129.039993
2020-09-01	134.800003	130.529999	132.759995	134.179993	152470100.0	134.179993
2020-09-02	137.979996	127.000000	137.589996	131.399994	200119000.0	131.399994
2020-09-03	128.839996	120.500000	126.910004	120.879997	257599600.0	120.879997
2020-09-04	123.699997	110.889999	120.070000	120.959999	332607200.0	120.959999
2020-09-08	118.989998	112.680000	113.949997	112.820000	230220200.0	112.820000
2020-09-09	118.470001	115.260002	117.260002	118.190002	68463412.0	118.190002

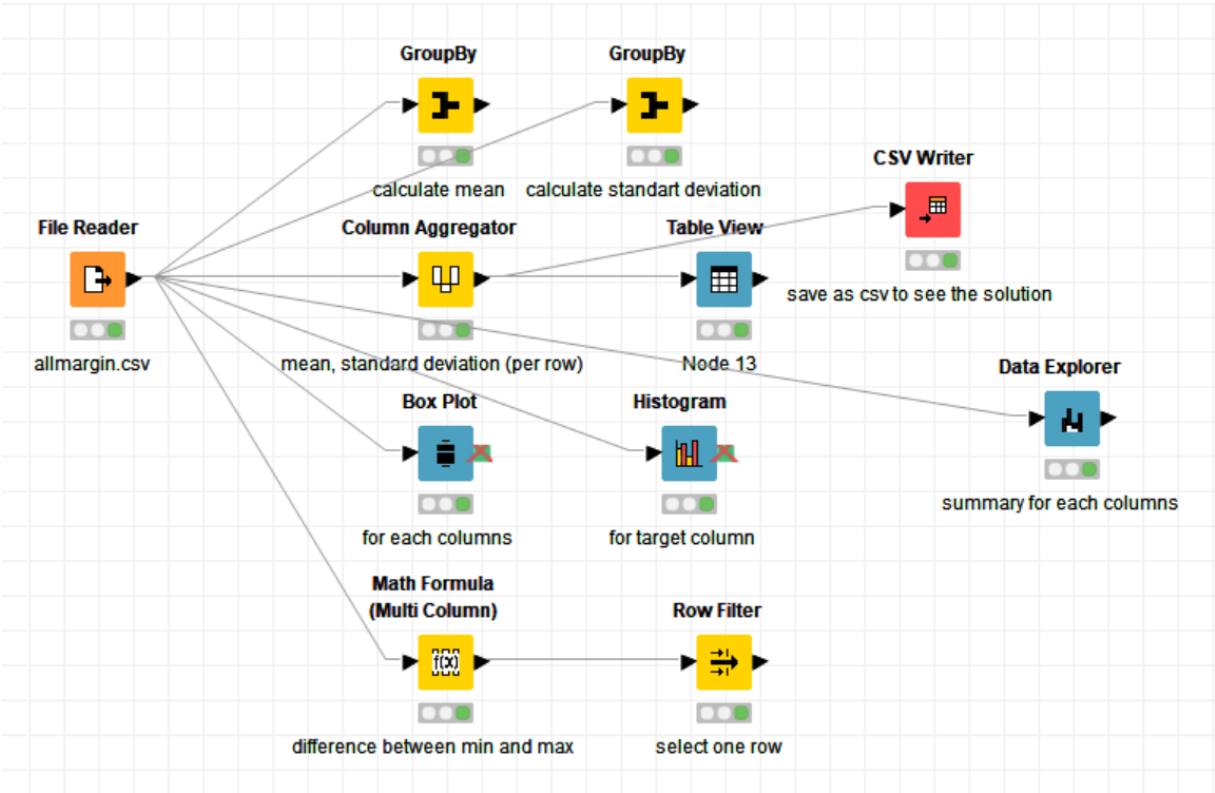
In each day, the dataset will be updated with highest stock price, lowest stock price, opening price of the stock, closing price of the stock, daily volume.

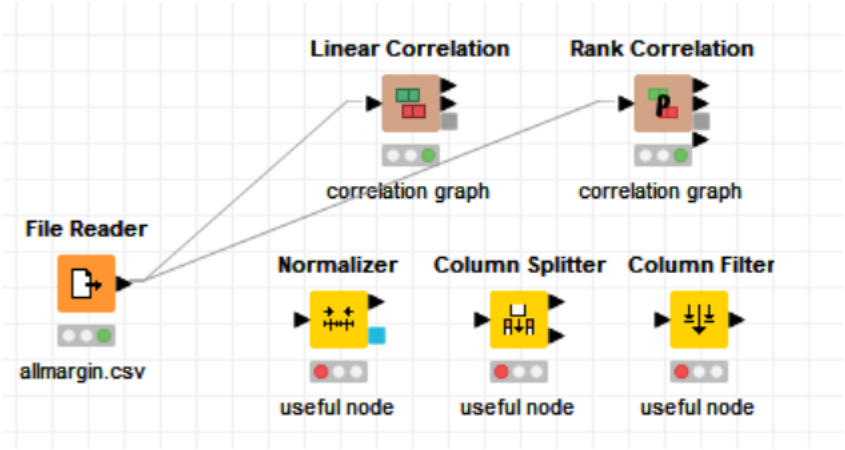
In summary, I captured Covid-19 disease information and yahoo stock price in each day of this week.

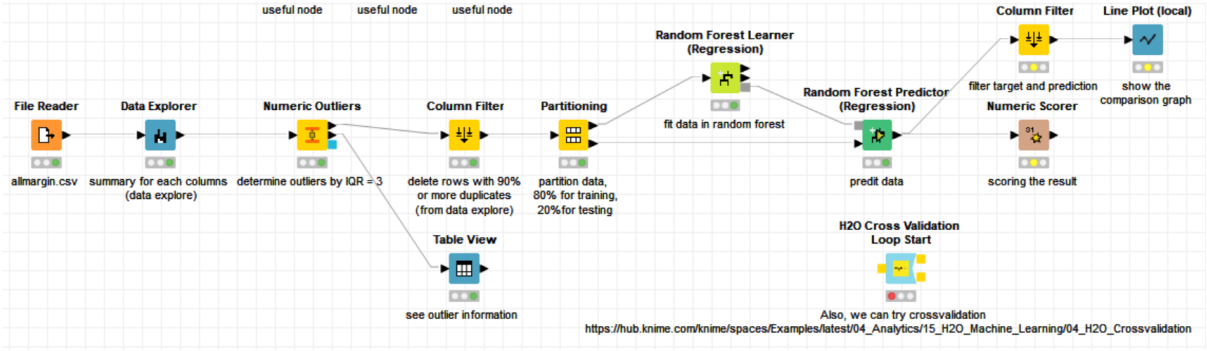
Question 2:

My workflow overview:

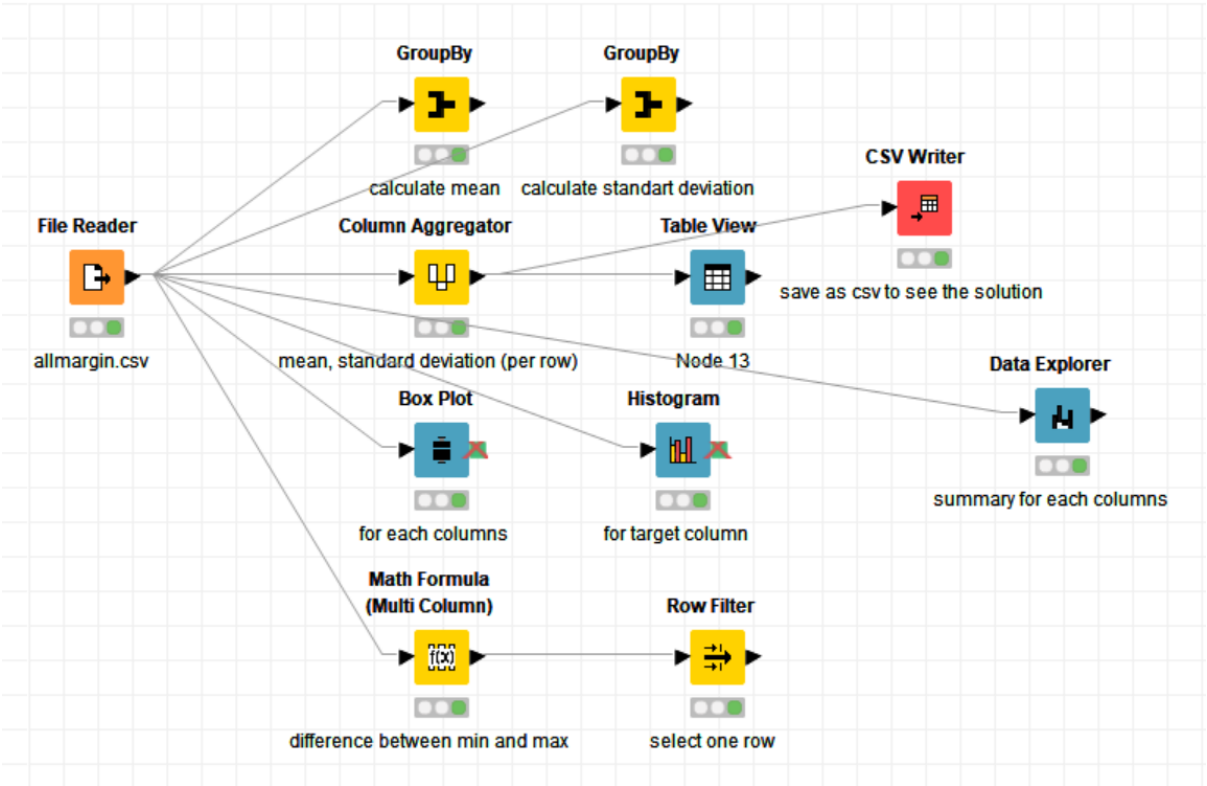




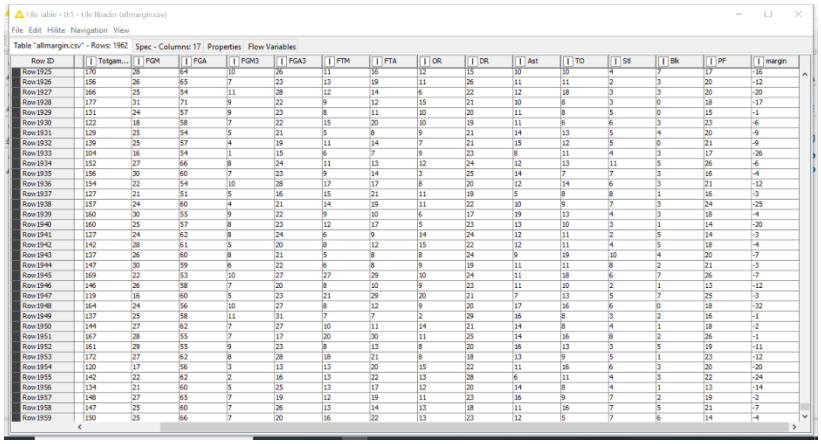




Explanation:



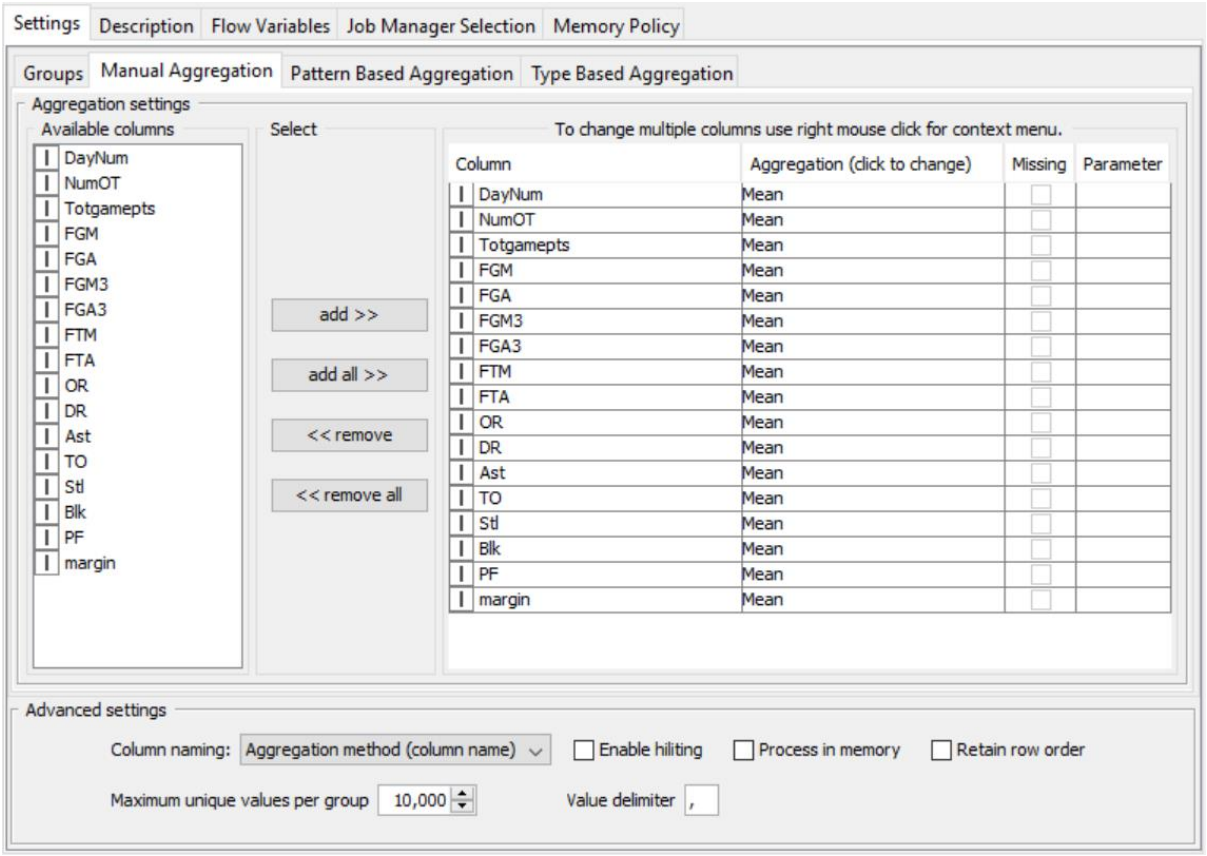
First, read data by file Reader node. I used the allmargin.csv dataset on OnQ for all processes below. By right click the node, click file table, then I got the following picture:



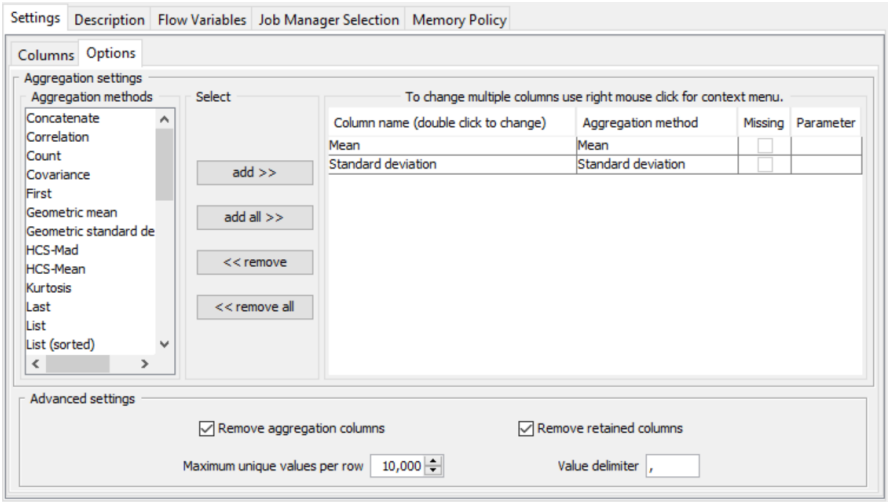
Then, we need to explore the dataset. Since there is nearly 2000 rows of data, I calculate mean and standard deviation, calculate min and max's differences, and use plots to visualize distributions.

First, for mean and standard deviation, I found several ways:

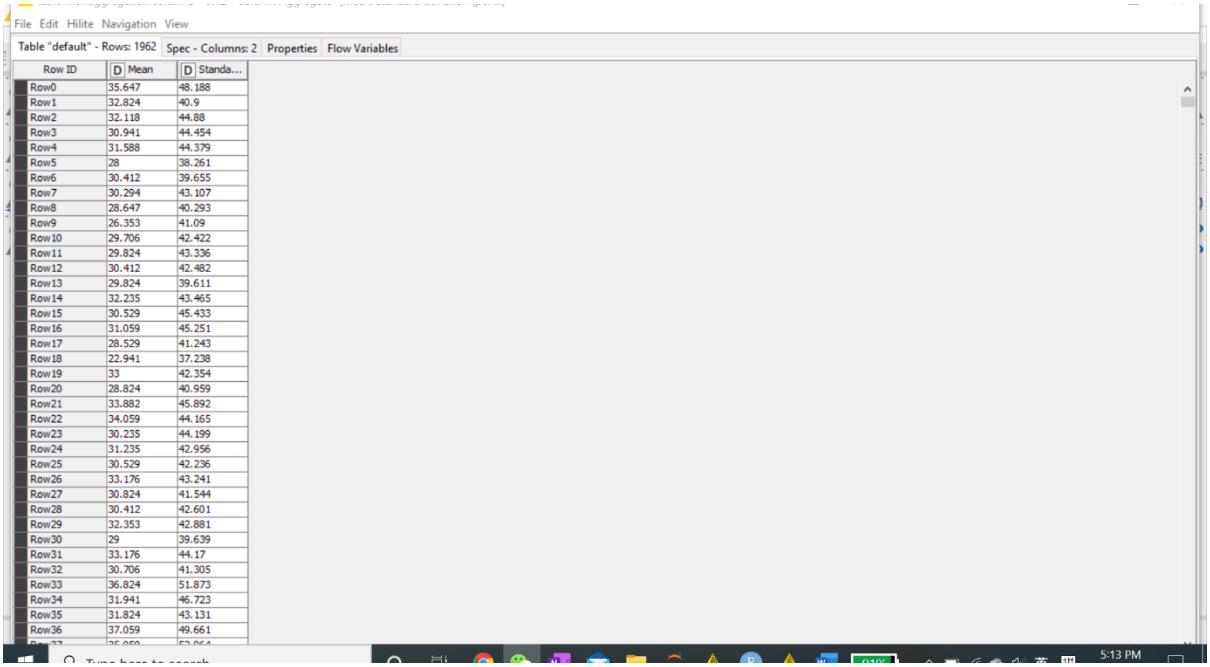
By groupby node → settings → manual aggregation → aggregation, we can calculate mean and standard deviation for each columns.



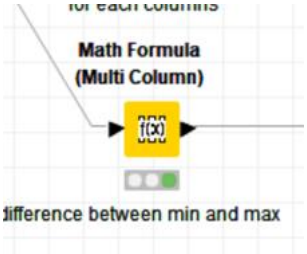
By column aggregator, setting → options →, we can select mean and standard deviation



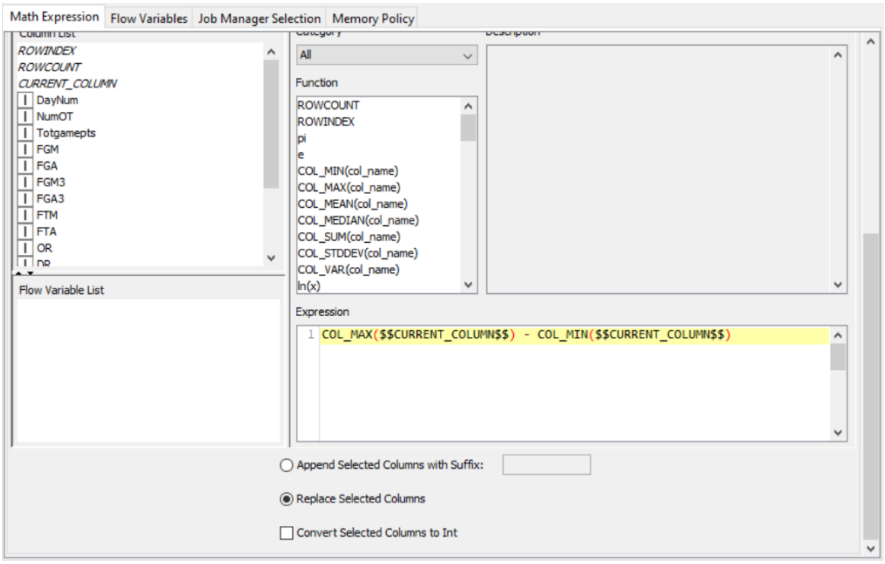
The result is the mean and standard deviations for each rows



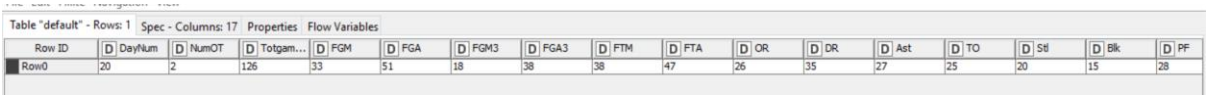
For calculate the difference between maximum and minimum value, I used math formula node:



Then setting the formula like this:

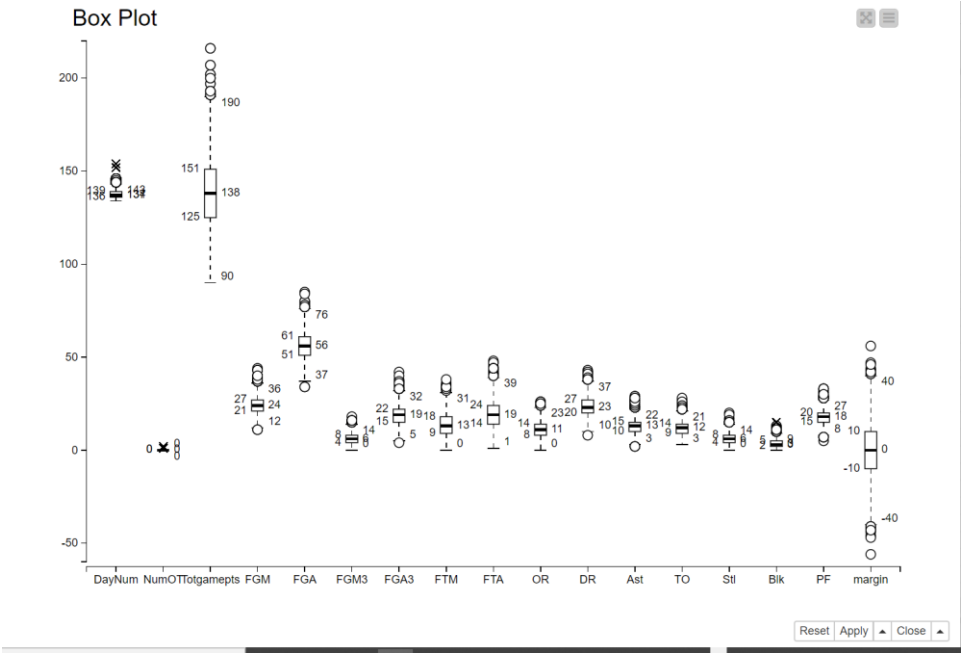


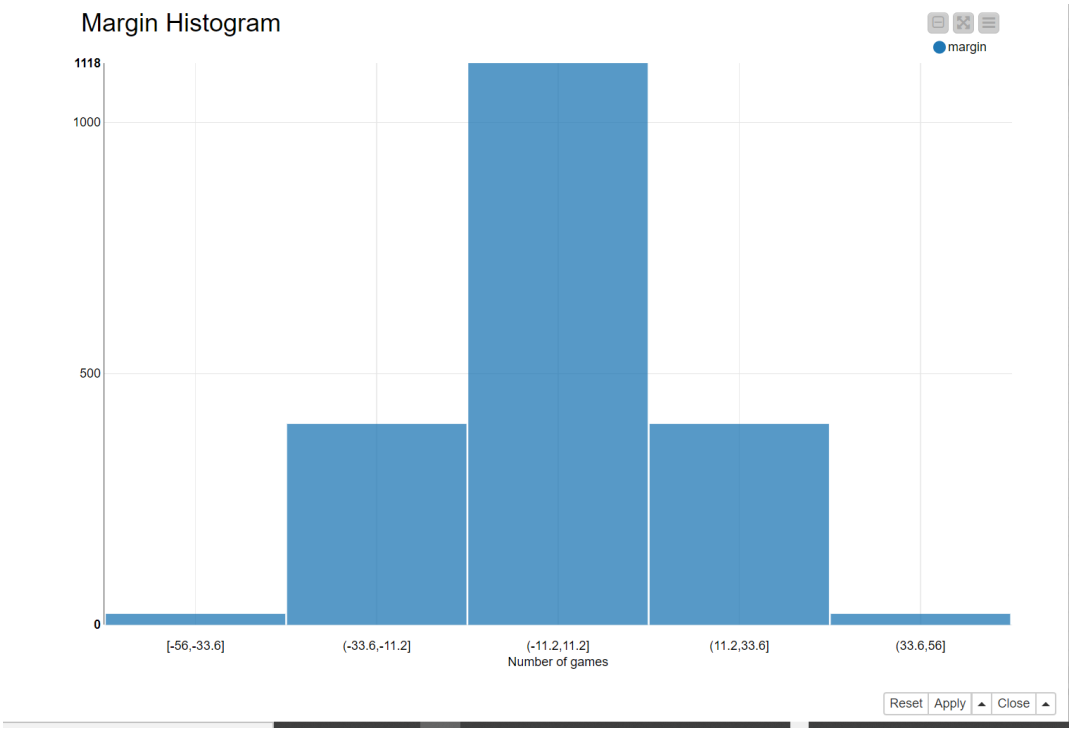
The result is difference between minimum and maximum value for each column. After filtering row, it shows like this:



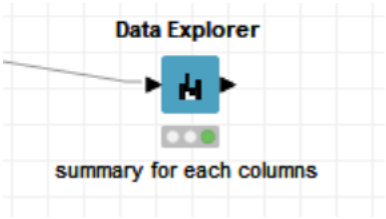
Row ID	DayNum	NumOT	Totgam...	FGM	FGA	FGM3	FGA3	FTM	FTA	OR	DR	Ast	TO	Stl	Blk	PF
Row0	20	2	126	33	51	18	38	38	47	26	35	27	25	20	15	28

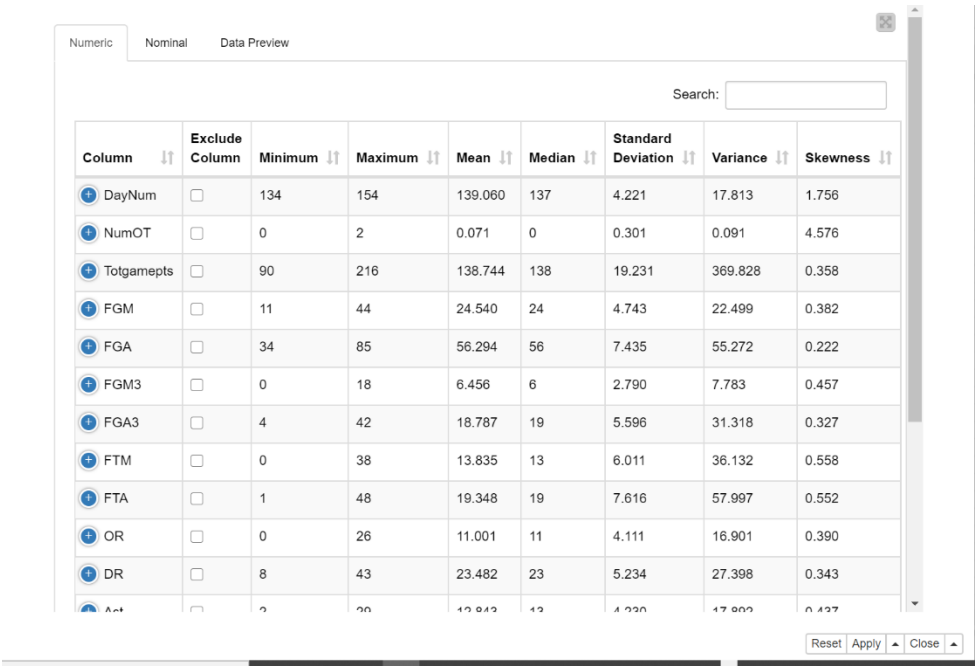
By box plot and histogram, boxplot visualize the outlier and range of data. Histogram counts data by groups. For example, I divided margin column into 5 groups with same intervals.





Compare with all steps above, data explorer is simpler and convenient.



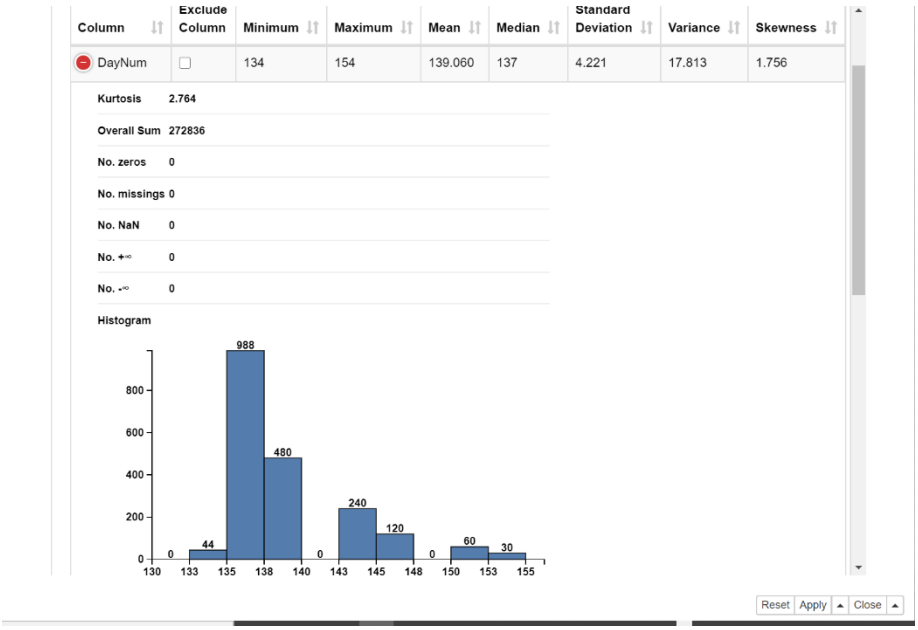


Numeric | Nominal | Data Preview

Search:

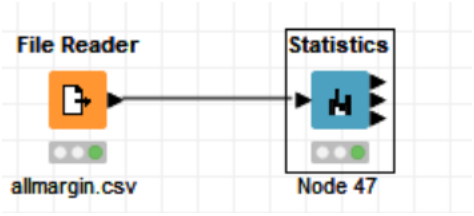
Column	Exclude Column	Minimum	Maximum	Mean	Median	Standard Deviation	Variance	Skewness
DayNum		134	154	139.060	137	4.221	17.813	1.756
NumOT		0	2	0.071	0	0.301	0.091	4.576
Totgamepts		90	216	138.744	138	19.231	369.828	0.358
FGM		11	44	24.540	24	4.743	22.499	0.382
FGA		34	85	56.294	56	7.435	55.272	0.222
FGM3		0	18	6.456	6	2.790	7.783	0.457
FGA3		4	42	18.787	19	5.596	31.318	0.327
FTM		0	38	13.835	13	6.011	36.132	0.558
FTA		1	48	19.348	19	7.616	57.997	0.552
OR		0	26	11.001	11	4.111	16.901	0.390
DR		8	43	23.482	23	5.234	27.398	0.343

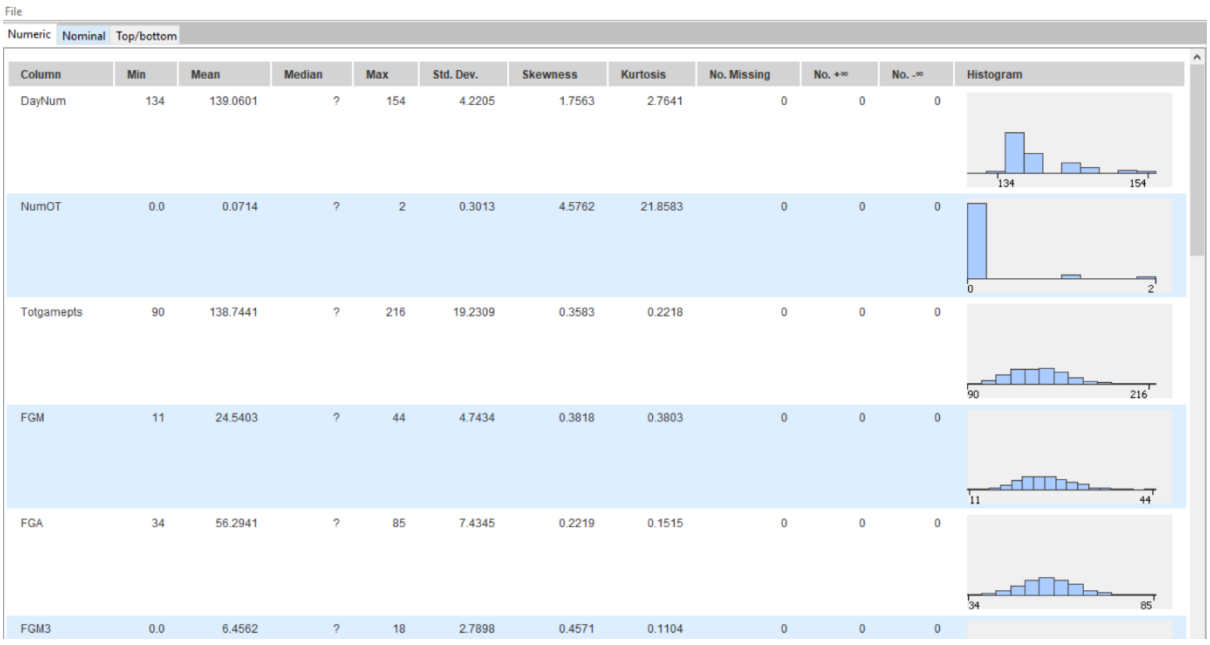
Reset | Apply | Close

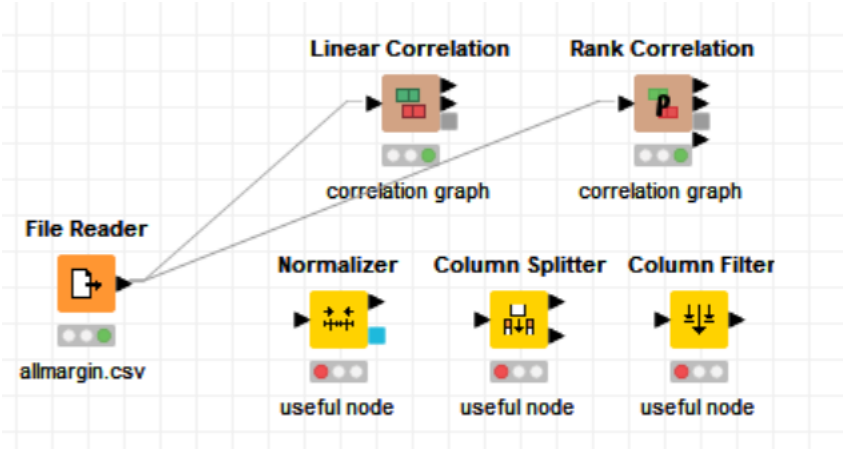


It includes minimum, maximum, mean, median, standard deviation, variance, skewness, total sum, number of missing values, and histogram for each columns.

Statistics node does the similar thing:

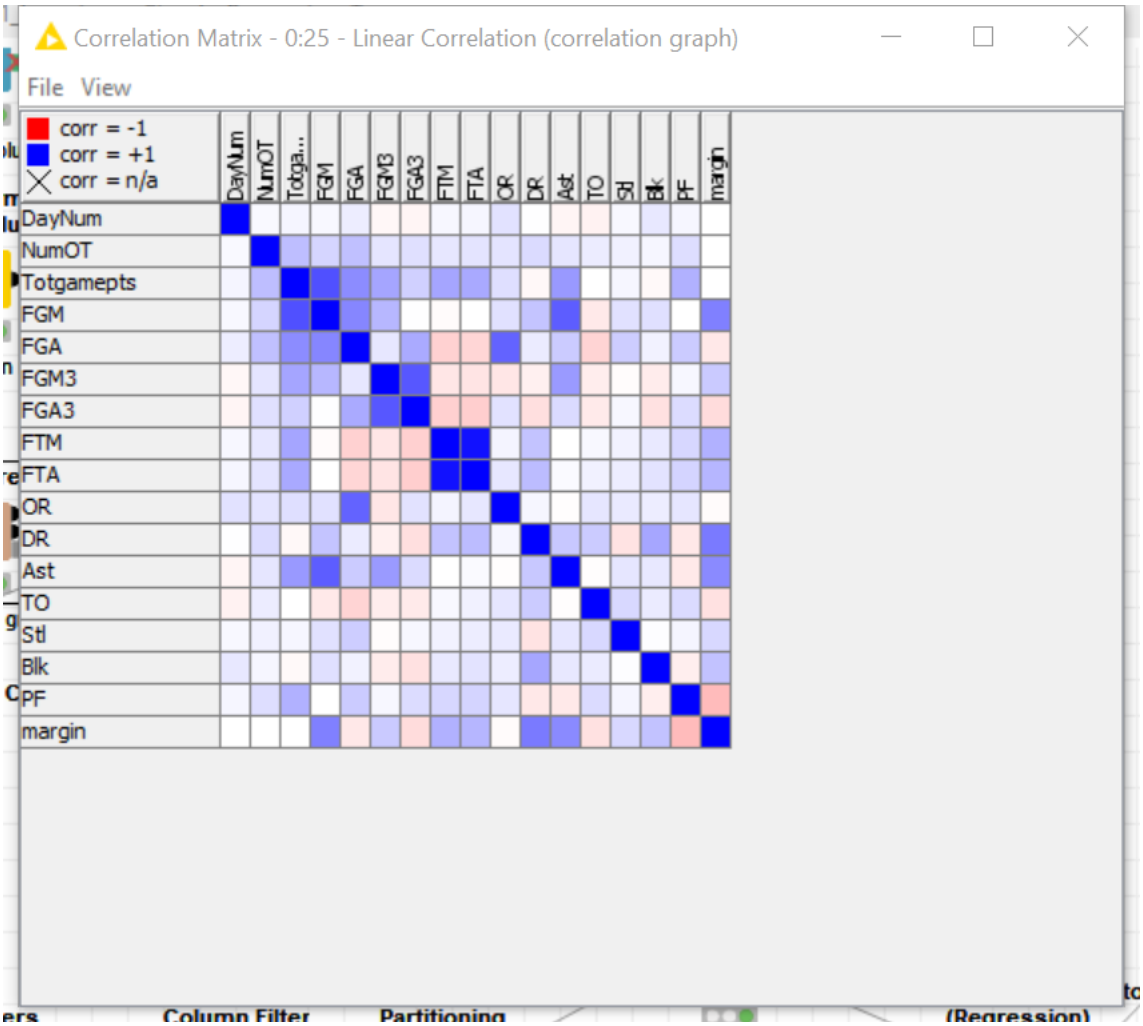






For correlation graphs, Linear Correlation and Rank Correlation nodes works well.

Correlation matrix graph:

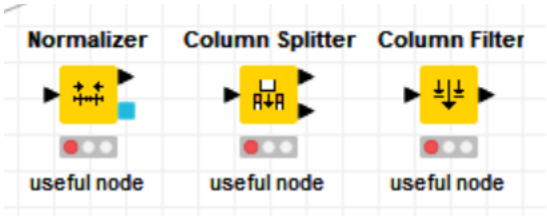


Correlation matrix:

Table "Correlation values" - Rows: 17 | Spec - Columns: 17 | Properties | Flow Variables

Row ID	DayNum	NumOT	Totgamepts	FGM	FGA	FGM3	FGA3	FTM	FTA	OR	DR
DayNum	1.0	0.022289039713431...	0.037723388174657...	0.026701471268 1...	0.070553965387 05...	-0.030482132606207...	-0.039659793164 37...	0.032350302903 00...	0.035966665177535...	0.114648517202 15...	0.0052
NumOT	0.022289039...	1.0	0.253104814902 31806	0.165695544238 6...	0.248564128662 05...	0.102612207705 69512	0.120298050459 8865	0.095748040293 02...	0.105641504155 45978	0.105366014879 53...	0.1398
Totgamepts	0.037723388...	0.253104814902 31...	1.0	0.686548614802625	0.449561109937 58...	0.354597131345 02735	0.183744724747 73967	0.351539066918 89...	0.337096189217 8963	0.126055577717 54...	-0.0285
FGM	0.026701471...	0.165695544238 63...	0.686548614802625	1.0	0.472586611244 3027	0.283291408396 40327	0.005787969272 391...	-0.017428793363 08...	-0.001898139725 684...	0.117323125863 08...	0.2316
FGA	0.070553965...	0.248564128662 05...	0.449561109937 58924	0.472586611244 3...	1.0	0.093545203136 62484	0.331905588126 66476	-0.179150865887 57...	-0.159224806053 0781	0.611995954421 4119	0.0769
FGM3	-0.030482132...	0.102612207705 69...	0.354597131345 02735	0.283291408396 4...	0.093545203136 662...	1.0	0.659625601476 329	-0.096628059461 72...	-0.104433642982 76454	-0.098636767451 84...	-0.0570
FGA3	-0.039659793...	0.120298050459 8865	0.183744724747 73967	0.005787969272 3...	0.331905588126 66...	0.659625601476 329	1.0	-0.185285990790 7635	-0.192246299629 34317	0.114066156325 03...	-0.1270
FTM	0.032350302...	0.095748040293 02...	0.351539066918 89886	-0.017428793363 0...	-0.179150865887 57...	-0.096628059461 72223	-0.185285990790 7635	1.0	0.931684735155 7316	0.042739647436 99...	0.2346
FTA	0.035966665...	0.105641504155 45...	0.337096189217 8963	-0.001898139725 6...	-0.159224806053 0781	-0.104433642982 76454	-0.192246299629 34...	0.931684735155 7316	1.0	0.095668555200 18...	0.2625
OR	0.114648517...	0.105366014879 53...	0.126055577717 54614	0.117323125863 0...	0.611995954421 4119	-0.098636767451 84462	0.114066156325 03825	0.042739647436 99...	0.095668555200 18475	1.0	0.0368
DR	0.005289857...	0.139872737413 13...	-0.028598387459 150...	0.231625814974 1...	0.076909084528 20...	-0.057063311842 449...	-0.127085842193 87...	0.234643380487 67...	0.262548116833 98116	0.036814659977 08...	1.0
Ast	-0.039373705...	0.096452562243 75...	0.401347788736 82386	0.635855355800 7...	0.203165414918 63...	0.394619621674 0922	0.139408498430 5027	-7.761713620573 94...	0.019256075604 35402	-0.008558223840 42...	0.2146
TO	-0.052829005...	0.075443104359 92...	-5.664374180438741...	-0.085589729786 4...	-0.167974745959 73...	-0.069960054680 16296	-0.083246669899 01...	0.027744645961 29...	0.055408796432 02837	0.098552999988 12...	0.2053
Stl	0.026956511...	0.054137532945 72...	0.033591037049 796...	0.118608128769 9...	0.194588126118 2063	-0.013207554721 819...	0.032196806574 604...	0.052679177192 87...	0.071793287825 246	0.081423549483 40...	-0.1109
Blk	0.089106077...	0.035413274749 76...	-0.025106866526 017...	0.121822080464 3...	0.051641014088 71...	-0.072972277353 48508	-0.116873090663 25...	0.086472961694 99...	0.108065480937 06402	0.070632188801 14...	0.3493
PF	0.034331324...	0.128035566992 4662	0.305402365803 6405	6.843088612405 4...	0.204319278976 48...	0.032346267466 83729	0.137348688090 23004	0.156432324434 42...	0.168594839538 50798	0.096959702677 62...	-0.0895
margin	6.001205537...	-7.203569714214 34...	-1.653162280420 937...	0.498922199451 6...	-0.091568076019 36...	0.207751467563 4266	-0.137683124212 48...	0.306818366367 2652	0.286389212705 5006	-0.016395496934 10...	0.5202

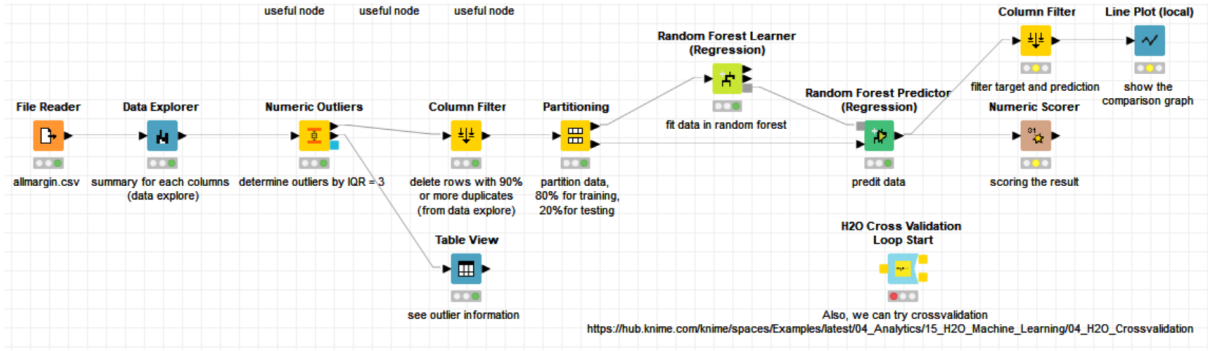
Other useful nodes:



These nodes are common for use.

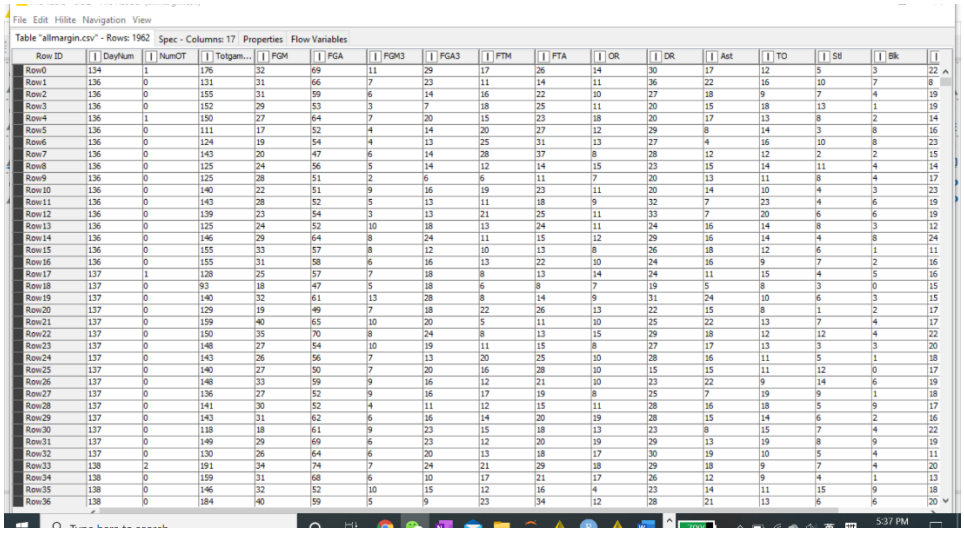
Prediction

I am trying to use KNIME for predictions. I used allmargin.csv as original dataset, then used outliers, partitioning, random forest Regressor, scoring and line plot.

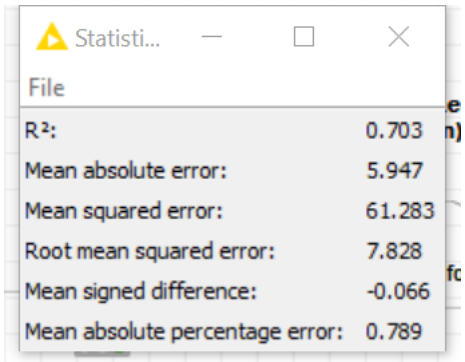


The comments are below each node.

Input:



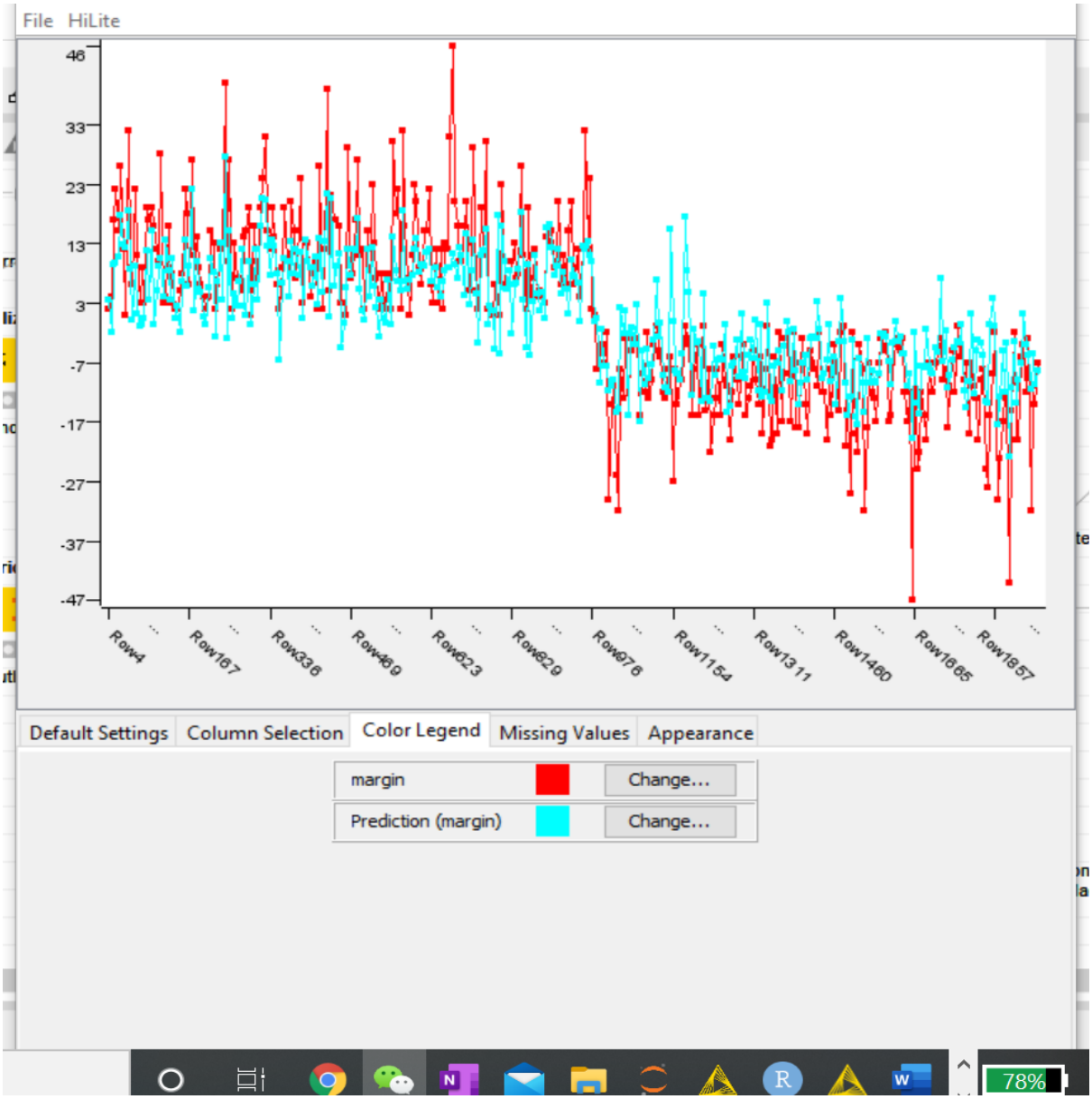
Outputs:



File

Table "Scores" - Rows: 6 | Spec - Column: 1 | Properties | Flow Variables

Column: 1	Column Type	Column Index	Color Handler	Size Handler	Shape Han...	Filter Handler	Lower Bound	Upper Bound
Prediction (margin)	Number (double)	0					-0.066	61.283


File HiLite
46
33
23
13
3
-7
-17
-27
-37
-47
Row4
Row167
Row336
Row469
Row623
Row829
Row976
Row1154
Row1311
Row1460
Row1685
Row1857
Default Settings
Column Selection
Color Legend
Missing Values
Appearance
margin
Change...
Prediction (margin)
Change...
78%