

## 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 x 23 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](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](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](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]:

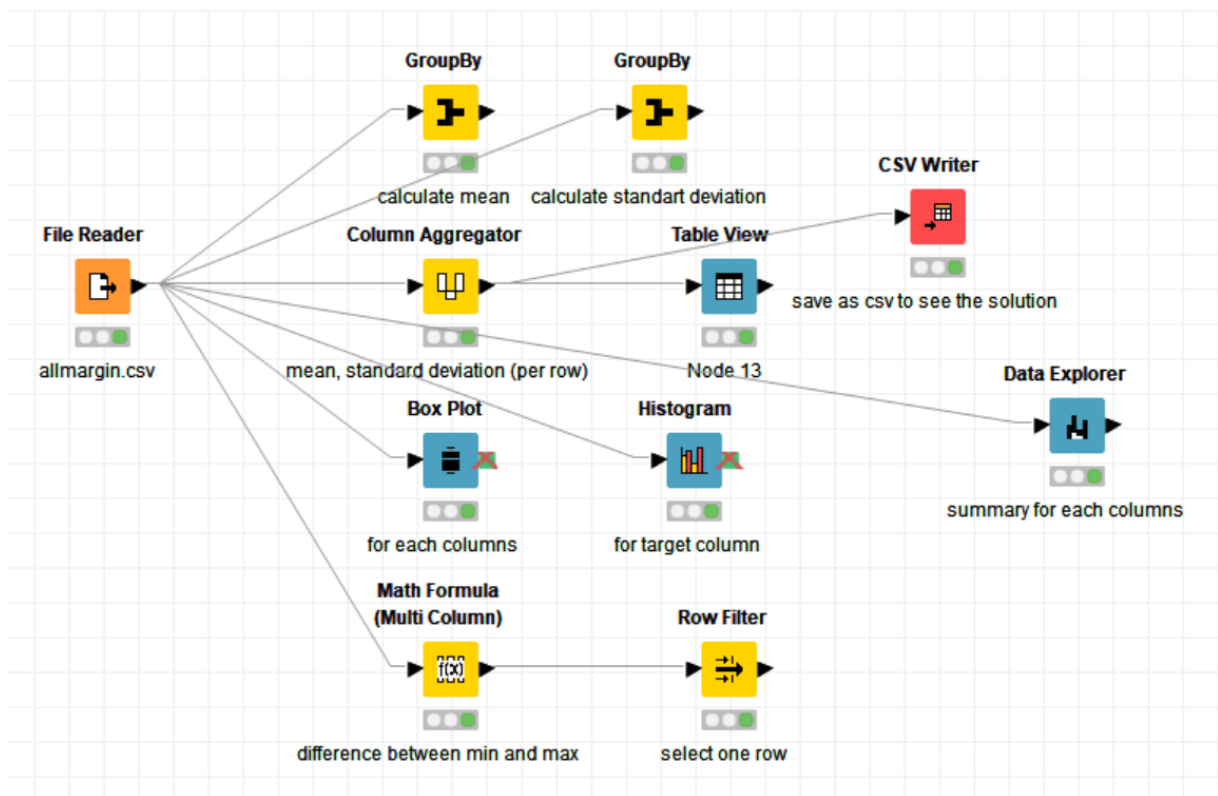
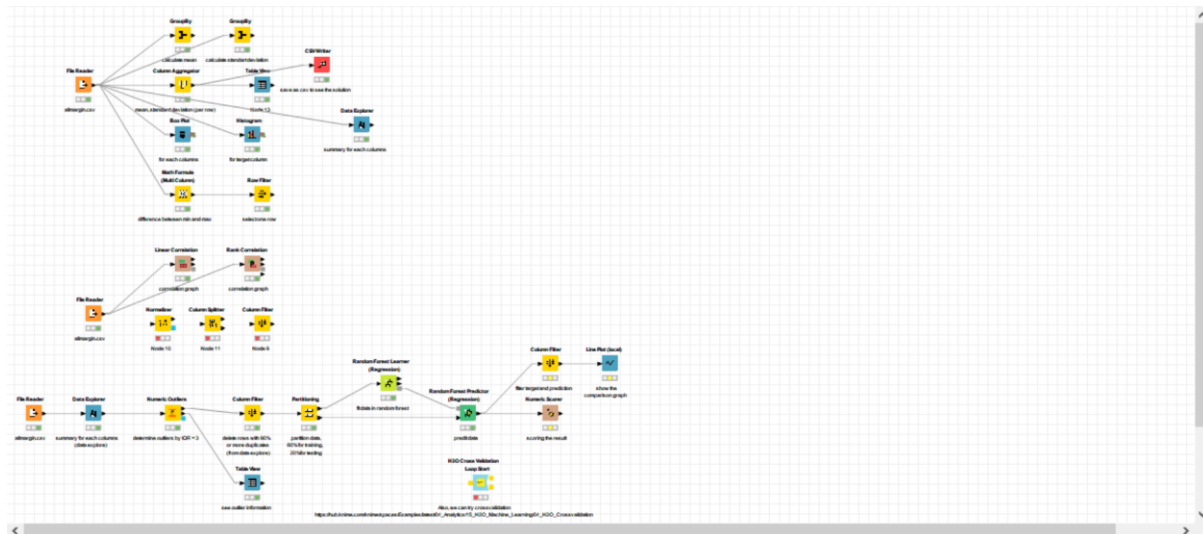
	High	Low	Open	Close	Volume	Adj Close
Date						
2020-08-24	128.785004	123.937500	128.697495	125.857498	345937600.0	125.857498
2020-08-25	125.180000	123.052498	124.697502	124.824997	211495600.0	124.824997
2020-08-26	126.992500	125.082497	126.180000	126.522499	163022400.0	126.522499
2020-08-27	127.485001	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

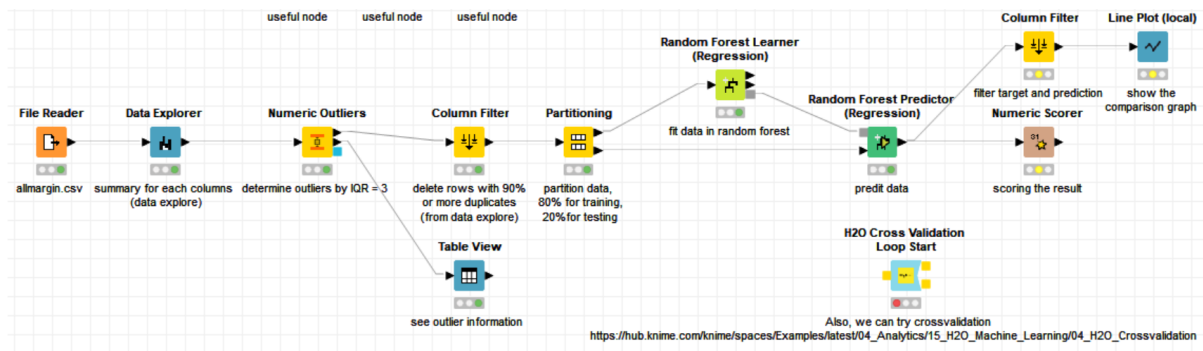
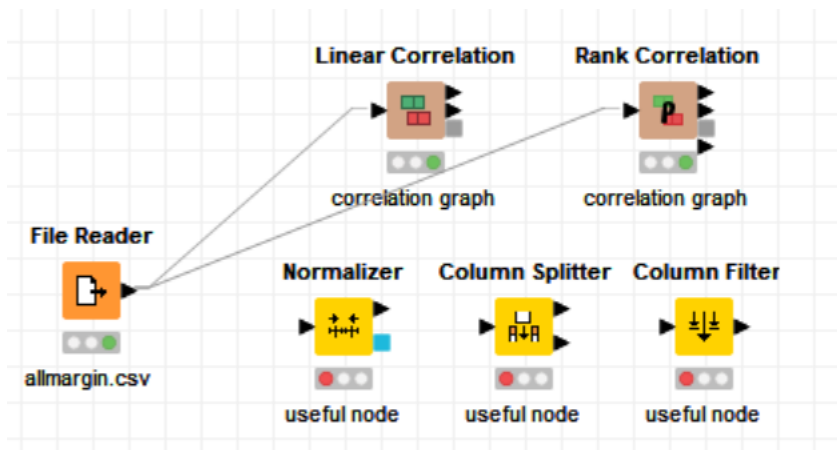
In each day, the dataset will be updated with highest stock price, lowest stock price, opening price of the stock, closing price of the stock, daily volume.

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

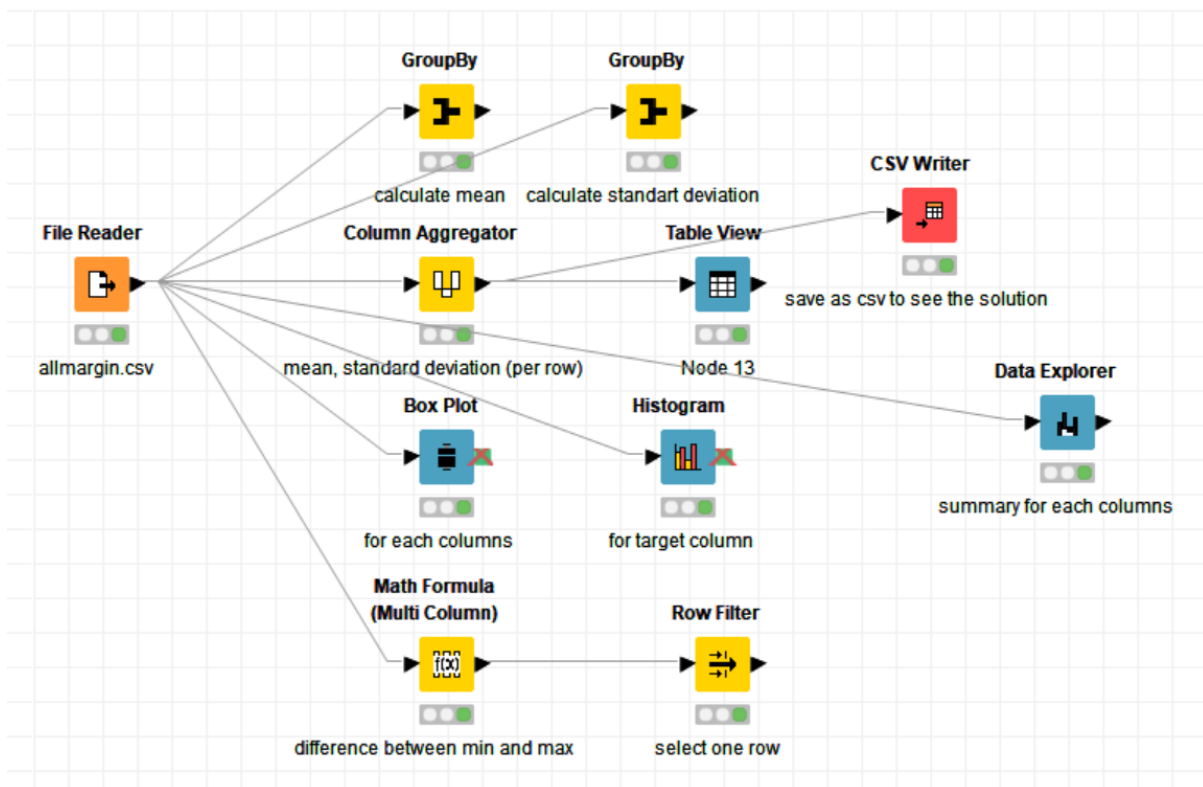
Question 2:

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:

Row ID	Totgame	FGM	FGA	FGM3	FGA3	FTM	FTA	OR	DR	Ast	TO	Stl	Blk	PF	margin
Row 1925	195	26	64	10	26	11	16	12	13	10	4	7	13	15	
Row 1926	156	26	65	7	23	13	19	11	26	11	11	2	3	20	-12
Row 1927	196	25	64	11	28	12	14	8	22	12	18	0	0	20	-20
Row 1928	177	31	71	9	23	8	12	15	21	10	8	0	0	18	-17
Row 1929	131	24	57	9	23	8	11	10	20	11	8	0	0	15	-1
Row 1930	122	18	58	7	22	15	30	10	19	11	6	6	3	23	-6
Row 1931	129	25	64	5	21	5	8	9	21	14	13	5	4	20	-9
Row 1932	139	25	57	4	19	11	14	7	21	15	12	5	0	21	-9
Row 1933	104	16	54	1	15	6	7	9	23	8	11	4	3	17	-28
Row 1934	152	27	66	8	24	11	13	12	24	12	13	11	5	26	-6
Row 1935	196	30	60	7	23	9	14	3	25	14	7	7	3	16	-4
Row 1936	154	22	54	10	28	17	17	8	20	12	14	6	3	21	-12
Row 1937	127	21	51	5	16	15	21	11	19	5	8	8	1	16	-3
Row 1938	137	24	60	4	21	14	19	11	22	10	9	7	3	24	-25
Row 1939	160	30	65	9	23	9	10	6	17	19	13	4	3	18	-4
Row 1940	160	25	57	8	23	12	17	5	23	13	10	3	1	14	-20
Row 1941	127	24	62	8	24	6	9	14	24	12	11	2	5	14	-3
Row 1942	142	28	61	5	20	8	12	15	22	12	11	4	5	18	-4
Row 1943	137	26	60	8	21	5	8	8	24	9	19	10	4	20	-7
Row 1944	147	30	69	6	32	8	9	19	11	11	8	2	1	21	-3
Row 1945	169	22	53	10	27	27	29	10	24	11	18	6	7	26	-7
Row 1946	146	26	58	7	20	8	10	9	23	11	10	2	1	13	-12
Row 1947	119	16	60	5	23	21	29	20	21	7	13	5	7	25	-3
Row 1948	164	24	56	10	27	8	12	9	20	17	16	6	0	18	-32
Row 1949	137	25	58	11	31	7	7	12	29	16	8	3	2	16	-1
Row 1950	144	27	62	7	27	10	11	14	21	14	8	4	1	18	-2
Row 1951	167	28	55	7	17	20	30	11	25	14	16	8	2	26	-1
Row 1952	161	29	55	9	23	8	13	8	20	16	13	3	5	19	-11
Row 1953	172	27	62	8	28	18	21	8	18	13	9	5	1	23	-12
Row 1954	120	17	56	3	13	13	20	15	22	11	16	6	3	20	-20
Row 1955	142	32	62	2	16	13	22	13	28	6	11	4	3	22	-24
Row 1956	134	21	60	5	25	13	17	12	20	14	8	4	1	13	-14
Row 1957	148	27	65	7	19	12	19	11	23	16	9	7	2	19	-2
Row 1958	147	25	60	7	26	13	14	13	18	11	16	7	5	21	-7
Row 1959	150	25	66	7	20	16	22	13	23	12	5	7	6	14	-4

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.

**Settings** | Description | Flow Variables | Job Manager Selection | Memory Policy

**Groups** | Manual Aggregation | Pattern Based Aggregation | Type Based Aggregation

**Aggregation settings**

Available columns: DayNum, NumOT, Totgamepts, FGM, FGA, FGM3, FGA3, FTM, FTA, OR, DR, Ast, TO, Stl, Blk, PF, margin

Select: add >>, add all >>, << remove, << remove all

To change multiple columns use right mouse click for context menu.

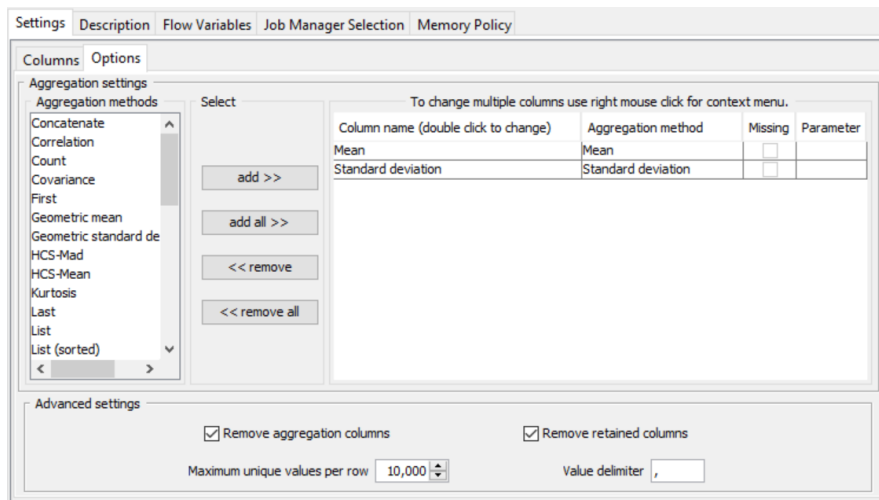
Column	Aggregation (click to change)	Missing	Parameter
DayNum	Mean	<input type="checkbox"/>	
NumOT	Mean	<input type="checkbox"/>	
Totgamepts	Mean	<input type="checkbox"/>	
FGM	Mean	<input type="checkbox"/>	
FGA	Mean	<input type="checkbox"/>	
FGM3	Mean	<input type="checkbox"/>	
FGA3	Mean	<input type="checkbox"/>	
FTM	Mean	<input type="checkbox"/>	
FTA	Mean	<input type="checkbox"/>	
OR	Mean	<input type="checkbox"/>	
DR	Mean	<input type="checkbox"/>	
Ast	Mean	<input type="checkbox"/>	
TO	Mean	<input type="checkbox"/>	
Stl	Mean	<input type="checkbox"/>	
Blk	Mean	<input type="checkbox"/>	
PF	Mean	<input type="checkbox"/>	
margin	Mean	<input type="checkbox"/>	

**Advanced settings**

Column naming: Aggregation method (column name) | Enable hiliting | Process in memory | Retain row order

Maximum unique values per group: 10,000 | Value delimiter: ,

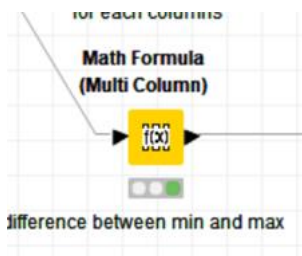
By column aggregator, setting → options →, we can select mean and standard deviation



The result is the mean and standard deviations for each rows

Row ID	D Mean	D Standa...
Row0	35.647	48.188
Row1	32.824	40.9
Row2	32.118	44.88
Row3	30.941	44.454
Row4	31.588	44.379
Row5	28	38.261
Row6	30.412	39.655
Row7	30.294	43.107
Row8	28.647	40.293
Row9	26.353	41.09
Row10	29.706	42.422
Row11	29.824	43.336
Row12	30.412	42.482
Row13	29.824	39.611
Row14	32.235	43.465
Row15	30.529	45.433
Row16	31.059	45.251
Row17	28.529	41.243
Row18	22.941	37.238
Row19	33	42.354
Row20	28.824	40.959
Row21	33.882	45.892
Row22	34.059	44.165
Row23	30.235	44.199
Row24	31.235	42.956
Row25	30.529	42.236
Row26	33.176	43.241
Row27	30.824	41.544
Row28	30.412	42.601
Row29	32.353	42.881
Row30	29	39.639
Row31	33.176	44.17
Row32	30.706	41.305
Row33	36.824	51.873
Row34	31.941	46.723
Row35	31.824	43.131
Row36	37.059	49.661
Row37	36.059	43.064

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

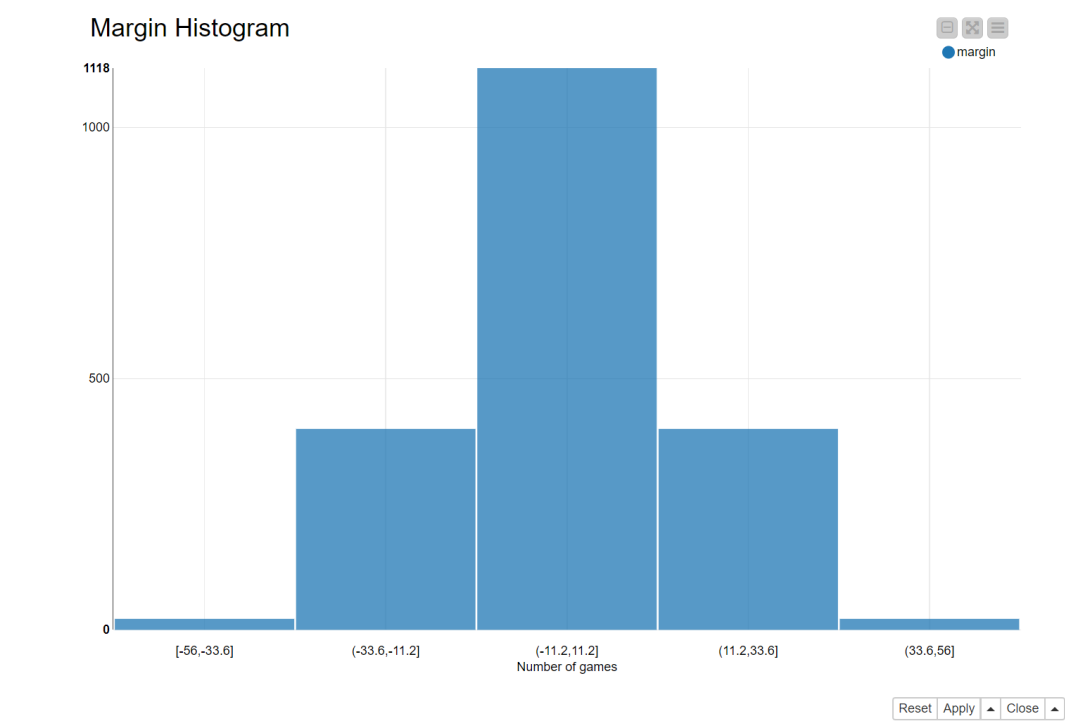


The screenshot shows the Tableau interface with the 'Columns' shelf. The 'Columns' shelf contains the expression `COL_MAX($CURRENT_COLUMN$) - COL_MIN($CURRENT_COLUMN$)`. The 'Function' list on the left shows `COL_MAX` and `COL_MIN` selected. The 'Expression' field on the right shows the formula `COL_MAX($CURRENT_COLUMN$) - COL_MIN($CURRENT_COLUMN$)`.

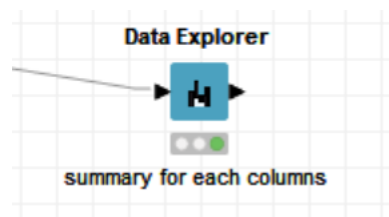
Row ID	DayNum	NumOT	TotalGame...	D_FGM	D_FGA	D_FGM3	D_FGA3	D_FTM	D_FTA	D_OR	D_DR	D_Ast	D_TO	D_Stl	D_Blk	D_PF
Row0	20	2	126	33	51	18	38	38	47	26	35	27	25	20	15	28

### Box Plot

Box plot showing the distribution of various basketball statistics. The y-axis represents the value of the statistic, ranging from -50 to 200. The x-axis lists the statistics: DayNum, NumOT, Tot, games, pts, FGM, FGA, FGM3, FGA3, FTM, FTA, OR, DR, Ast, TO, Stl, Blk, PF, and margin. Each statistic has a box plot with a median line, a box representing the interquartile range, and whiskers extending to the range of the data. Outliers are shown as open circles. The 'margin' statistic has a very wide range, with outliers at -40 and 40. The 'games' statistic has a very high outlier at 190. The 'PF' statistic has a very low outlier at -10. The 'margin' statistic has a very low outlier at -40 and a high outlier at 40.



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



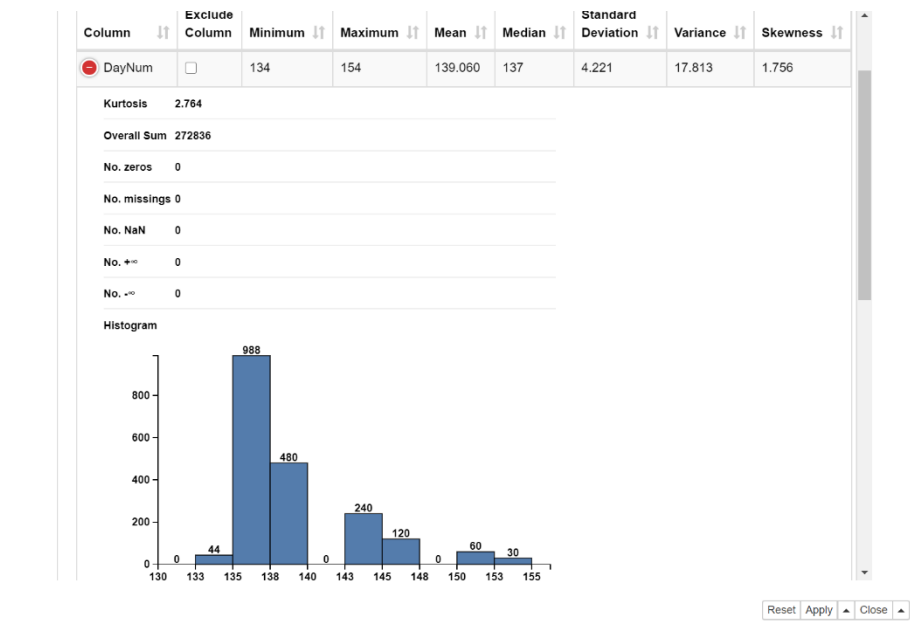
Numeric Nominal Data Preview

Search:

Column	Exclude Column	Minimum	Maximum	Mean	Median	Standard Deviation	Variance	Skewness
DayNum	<input type="checkbox"/>	134	154	139.060	137	4.221	17.813	1.756
NumOT	<input type="checkbox"/>	0	2	0.071	0	0.301	0.091	4.576
Totgamepts	<input type="checkbox"/>	90	216	138.744	138	19.231	369.828	0.358
FGM	<input type="checkbox"/>	11	44	24.540	24	4.743	22.499	0.382
FGA	<input type="checkbox"/>	34	85	56.294	56	7.435	55.272	0.222
FGM3	<input type="checkbox"/>	0	18	6.456	6	2.790	7.783	0.457
FGA3	<input type="checkbox"/>	4	42	18.787	19	5.596	31.318	0.327
FTM	<input type="checkbox"/>	0	38	13.835	13	6.011	36.132	0.558
FTA	<input type="checkbox"/>	1	48	19.348	19	7.616	57.997	0.552
OR	<input type="checkbox"/>	0	26	11.001	11	4.111	16.901	0.390
DR	<input type="checkbox"/>	8	43	23.482	23	5.234	27.398	0.343
Ass	<input type="checkbox"/>	2	26	12.843	13	4.230	17.892	0.437

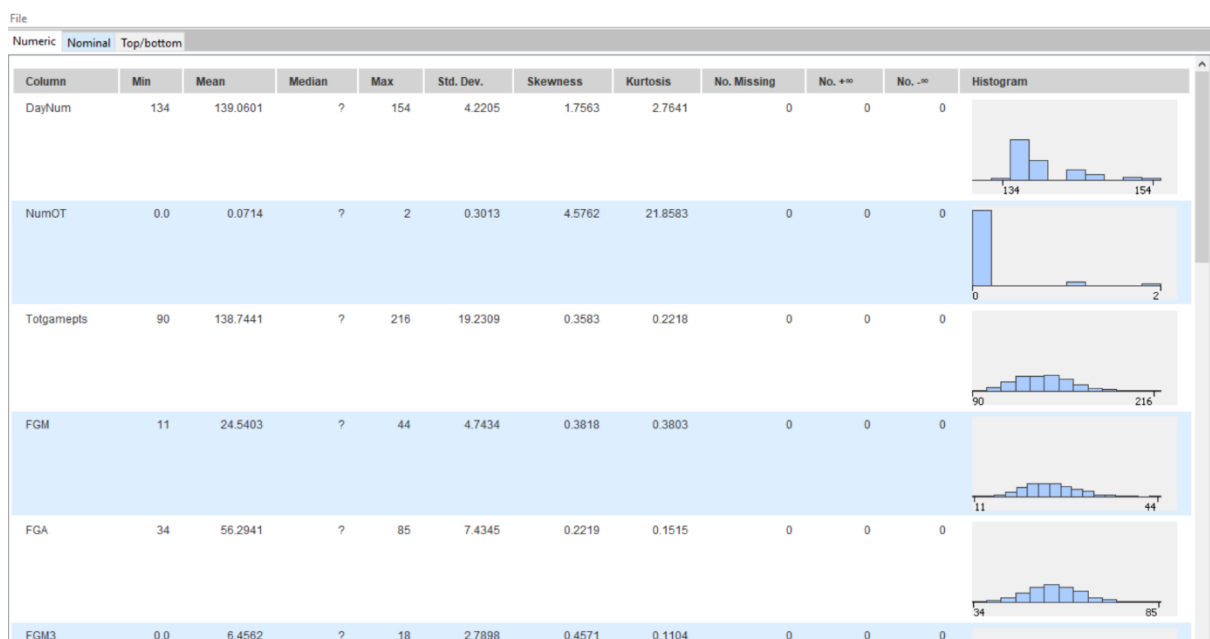
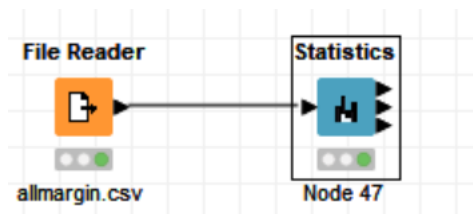
Reset Apply Close





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:

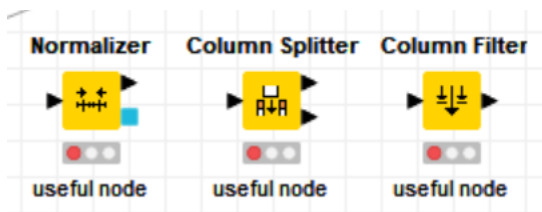




Correlation matrix:

Row ID	D Daylum	D NumOT	D Totgamepts	D FGM	D FGA	D FGM3	D FGA3	D FTM	D FTA	D OR	D DR
Daylum	1.0	0.02228903971343...	0.037723388174657...	0.0267014712681...	0.07055396538705...	-0.030482132606207...	-0.03965979316437...	0.03235030290300...	0.035966665177535...	0.11464851720215...	0.00523
NumOT	0.02228903...	1.0	0.25310481490231806	0.1656955442386...	0.24856412866205...	0.10261220770569512	0.1202980504509865	0.09574804029302...	0.10564150415545978	0.10536601487953...	0.1398
Totgamepts	0.03772338...	0.25310481490231...	1.0	0.686548614802625	0.44956110993758...	0.35459713134502735	0.18374472474773967	0.35153906691889...	0.3370961892178963	0.12605557771754...	-0.0285
FGM	0.02670147...	0.16569554423863...	0.686548614802625	1.0	0.4725866112443027	0.28329140839640327	0.005787969272391...	-0.01742879336308...	-0.001898139725684...	0.11732312586308...	0.2316
FGA	0.07055396...	0.24856412866205...	0.44956110993758924	0.4725866112443...	1.0	0.09354520313662484	0.33190558812666476	-0.17915086588757...	-0.1592248060530781	0.6119959544214119	0.0769
FGM3	-0.0304821...	0.10261220770569...	0.35459713134502735	0.2832914083964...	0.09354520313662...	1.0	0.659625601476329	-0.09662805946172...	-0.10443364298276454	-0.09863676745184...	-0.0570
FGA3	-0.0396597...	0.1202980504509865	0.18374472474773967	0.0057879692723...	0.33190558812666...	0.659625601476329	1.0	-0.1852899907907635	-0.19224629962934317	0.11406615632503...	-0.1270
FTM	0.03235030...	0.09574804029302...	0.35153906691889886	-0.0174287933630...	-0.17915086588757...	-0.09662805946172223	-0.1852899907907635	1.0	0.9316847351557316	0.04273964743699...	0.2346
FTA	0.03596666...	0.10564150415545...	0.3370961892178963	-0.0018981397256...	-0.1592248060530781	-0.10443364298276454	-0.19224629962934...	0.9316847351557316	1.0	0.09566855520018...	0.2675
OR	0.11464851...	0.10536601487953...	0.12605557771754614	0.1173231258630...	0.6119959544214119	-0.0986367674518462	0.11406615632503825	0.04273964743699...	0.09566855520018475	1.0	0.368
DR	0.00528985...	0.13987273741313...	-0.028598387459150...	0.2316258149741...	0.07690908452820...	-0.057063311842448...	-0.12708584219387...	0.23464338048767...	0.26254811683398116	0.3681465997708...	1.0
Ast	-0.0393737...	0.09645256224375...	0.40134778873682386	0.6358553580007...	0.20316541491863...	0.3946196216740922	0.1394084984305027	-7.76171362057794...	0.01925607560435402	-0.00855822384042...	0.2146
TO	-0.0528290...	0.07544310435992...	-5.664374180438741...	-0.0855897297864...	-0.16797474595973...	-0.06996005468016296	-0.08324666899901...	0.02774464596129...	0.05540879643202837	0.09855299998812...	0.2053
SH	0.02695651...	0.0541375294572...	0.033591037049796...	0.1186081287699...	0.1945881261182063	-0.013207554721819...	0.032196806574604...	0.05267917719287...	0.071793287825246	0.08142354948340...	-0.1106
Bk	0.08910607...	0.03541327474976...	-0.025106866526017...	0.1218220804643...	0.05164101408871...	-0.07297227735348508	-0.11687309066325...	0.08647296169499...	0.10806548093706402	0.07063218880114...	0.3493
PF	0.03433132...	0.1280355649924662	0.3054023658036405	6.8430886124054...	0.20431927897648...	0.03234626746683729	0.13734868809023004	0.15643232443442...	0.16859483953850798	0.09695970267762...	-0.0895
margin	6.00120553...	-7.20356971421434...	-1.653162280420937...	0.4989221994516...	-0.09156807601936...	0.2077514675634266	-0.13768312421248...	0.3068183663672652	0.2863892127055006	-0.01639549693410...	0.5202

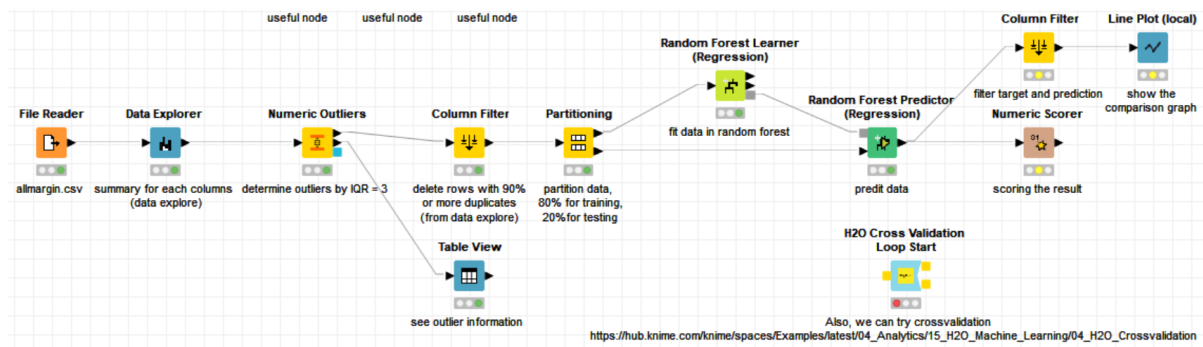
Other useful nodes:



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:

Table "allmargin.csv" - Rows: 1962																
Row ID	DayNum	NumOT	Totgam...	FGM	FGA	FGM3	FGA3	FTM	FTA	OR	DR	Ass	TO	SB	Bk	
Row0	134	1	176	32	89	11	29	17	26	14	30	17	12	5	3	22
Row1	136	0	131	31	66	7	23	11	14	11	36	22	16	10	7	8
Row2	136	0	155	31	99	6	14	16	22	10	27	18	9	7	4	19
Row3	136	0	152	29	53	3	7	18	25	11	20	15	18	13	1	19
Row4	136	1	150	27	64	7	20	15	23	18	20	17	13	8	2	14
Row5	136	0	111	17	52	4	14	20	27	12	29	8	14	3	8	16
Row6	136	0	124	19	54	4	13	25	31	13	27	4	16	10	8	23
Row7	136	0	143	20	47	6	14	28	37	8	28	12	12	2	2	15
Row8	136	0	125	24	56	5	14	12	14	15	23	15	14	11	4	14
Row9	136	0	125	28	51	2	6	6	11	7	20	13	11	8	4	17
Row10	136	0	140	22	51	9	16	19	23	11	20	14	10	4	3	23
Row11	136	0	143	28	52	5	13	11	18	9	32	7	23	4	6	19
Row12	136	0	139	23	54	3	13	21	25	11	33	7	20	6	6	19
Row13	136	0	125	24	52	10	19	13	24	11	24	16	14	8	3	12
Row14	136	0	146	29	64	8	24	11	15	12	29	16	14	4	8	24
Row15	136	0	155	33	57	8	12	10	13	8	26	18	12	6	1	11
Row16	136	0	155	31	58	6	16	13	22	10	24	16	9	7	2	16
Row17	137	1	128	25	57	7	18	8	13	14	24	11	15	4	5	16
Row18	137	0	93	18	47	5	18	6	8	7	19	5	8	3	0	15
Row19	137	0	140	32	61	13	28	8	14	9	31	24	10	6	3	15
Row20	137	0	129	19	49	7	18	22	26	13	22	15	8	1	2	17
Row21	137	0	159	40	65	10	20	5	11	10	25	22	13	7	4	17
Row22	137	0	150	35	70	8	24	8	13	15	29	18	12	12	4	22
Row23	137	0	148	27	54	10	19	11	15	8	27	17	13	3	3	20
Row24	137	0	143	26	56	7	13	20	23	10	28	16	11	5	1	18
Row25	137	0	140	27	50	7	20	16	28	10	15	15	11	12	0	17
Row26	137	0	148	33	59	9	16	12	21	10	23	22	9	14	6	19
Row27	137	0	136	27	52	9	16	17	19	8	25	7	19	9	1	18
Row28	137	0	141	30	52	4	11	12	15	11	28	16	18	5	9	17
Row29	137	0	143	31	62	6	16	14	20	19	28	15	14	6	2	16
Row30	137	0	118	18	61	9	23	15	18	13	23	8	15	7	4	22
Row31	137	0	149	29	69	6	23	12	20	19	29	13	19	8	9	19
Row32	137	0	130	26	64	6	20	13	18	17	30	19	10	5	4	11
Row33	138	2	191	34	74	7	24	21	29	18	29	18	9	7	4	20
Row34	138	0	159	31	68	6	10	17	21	17	26	12	9	4	1	13
Row35	138	0	146	32	52	10	15	12	16	4	23	14	11	15	9	18
Row36	138	0	184	40	59	5	9	23	34	12	28	21	13	6	6	20

Outputs:

Statisti...	
File	
R <sup>2</sup> :	0.703
Mean absolute error:	5.947
Mean squared error:	61.283
Root mean squared error:	7.828
Mean signed difference:	-0.066
Mean absolute percentage error:	0.789

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

