R-flow™

-환경데이터 분석,AI 전문가-

환경정보 융합 빅데이터 플랫폼 데이터셋 분석하기

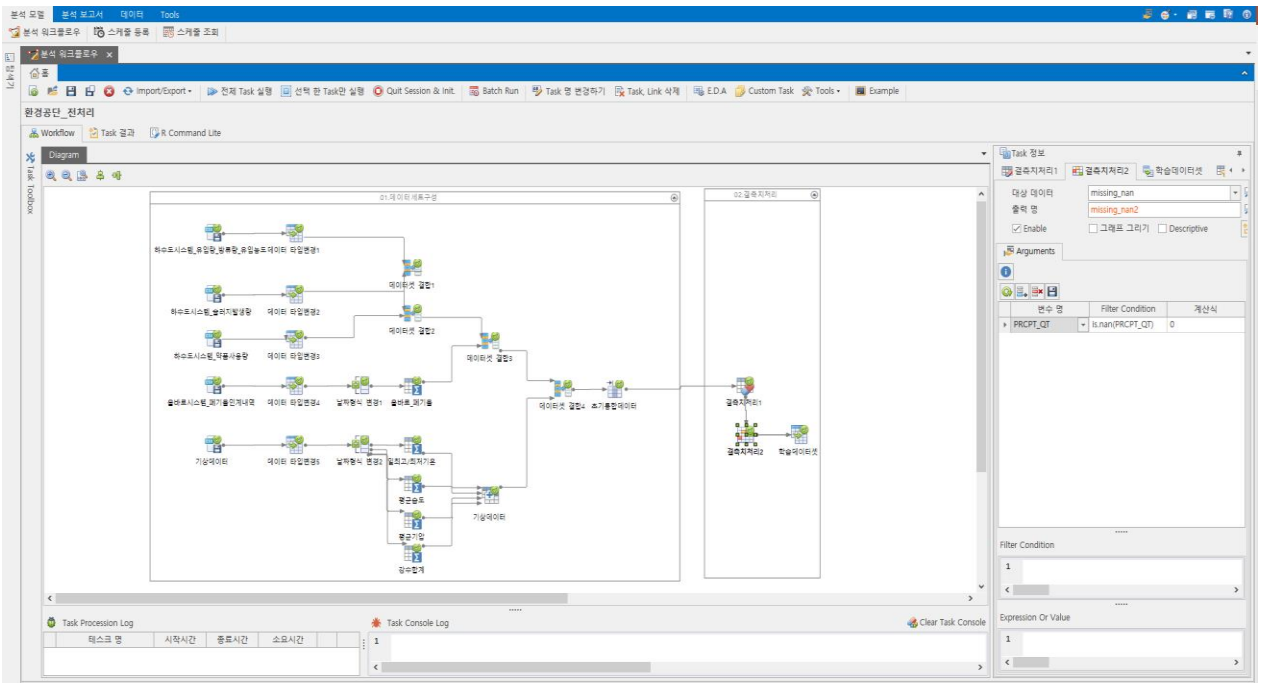
환경부

R-flow™

목 차

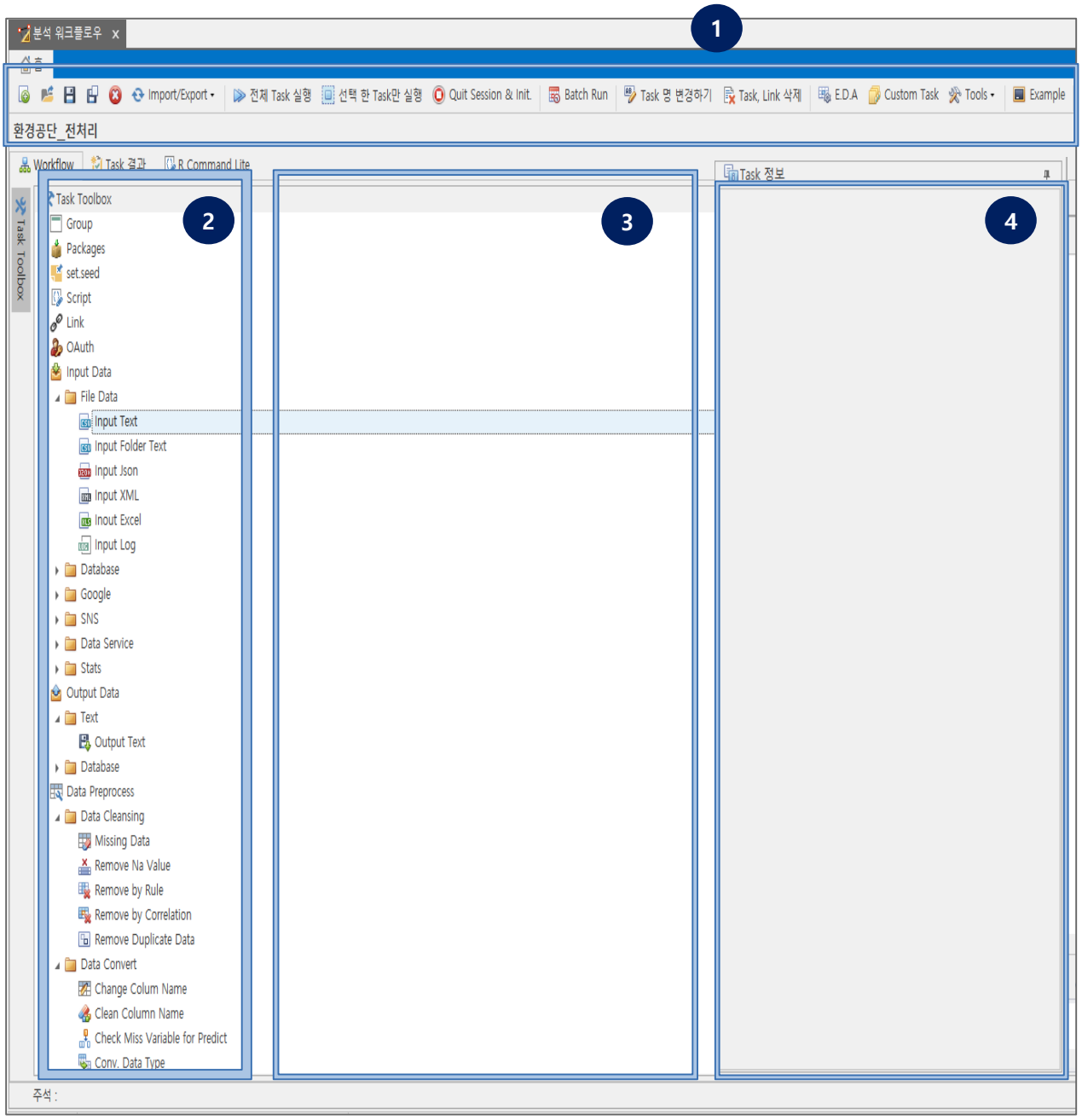
시스템 개요

- R AnalyticFlow는 통계 컴퓨팅을 위해 R 환경을 활용하는 데이터 분석 소프트웨어입니다.
- 직관적 인 사용자 인터페이스 외에도 R 전문가를위한 고급 기능을 제공합니다. 이러한 기능을 사용하면 서로 다른 수준의 숙련도를 가진 사용자간에 데이터 분석 프로세스를 공유 할 수 있습니다.



- R AnalyticFlow는 워크 플로우에서 데이터 분석 프로세스를 구성합니다. 시각화 된 프로세스는 마우스를 사용하여 간단하고 정확하게 재현 할 수 있습니다.
- 보다 편리하고 사용자 친화적 인 방식으로 선도적 인 데이터 분석 엔진의 성능을 극대화합니다.
- R AnalyticFlow는 실제 데이터를 분석하는 다양한 기능을 갖추고 있습니다. 데이터 읽기, 전처리, 그래프 작성, 통계 처리 및 예측 모델링과 같은 풍부한 응용 프로그램을 즉시 사용할 수 있습니다.
- R AnalyticFlow를 사용하면 옵션을 선택하고 결과를 미리 볼 수 있으므로 대화식으로 분석을 수행 할 수 있습니다. 사용자는 신속하고 정확하게 프로세스를 설명하고 편집하고 다른 사용자와 결과를 공유 할 수 있습니다.
- 또한 숙련 된 R 사용자의 프로그래밍 부담을 줄이고 세부 옵션을 지정하거나 자동 생성 스크립트를 수동으로 편집 할 수 있습니다.
- R AnalyticFlow는 풍부한 지원 기능을 갖추고 있습니다. 몇 가지 예를 들자면, 객체 브라우저는 분석 결과를 신속하게 확인하고, 처리 결과 저장 및 재사용을위한 객체 캐싱, 디버깅 기능 및 자동 백업 시스템을 제공합니다. 이러한 모든 기능이 분석을 강력하게 지원합니다.

1 분석 워크플로우 메인 화면



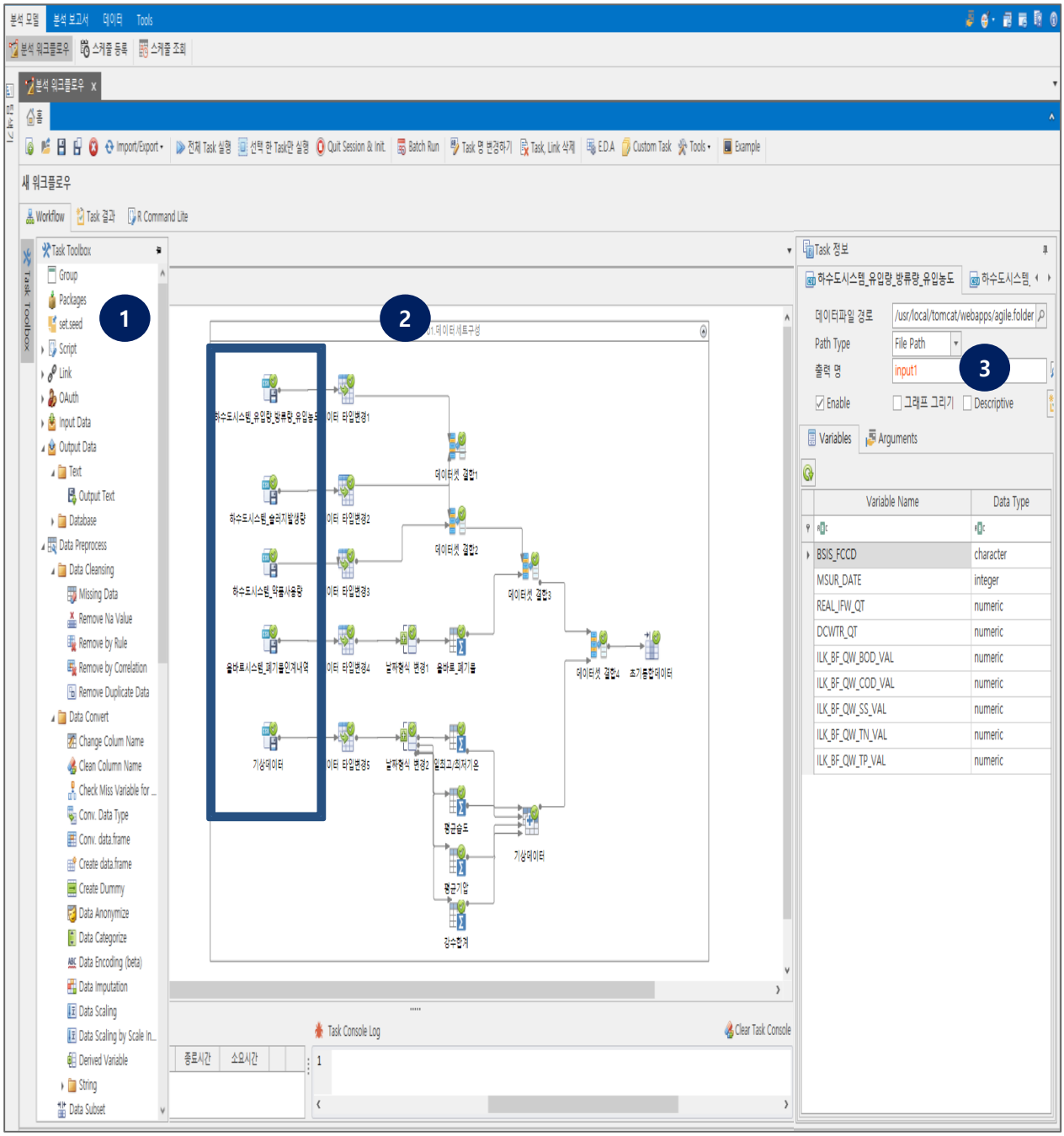
① 워크플로우에 대한 생성, 저장, 실행, EDA분석, 실행예제 등 다양한 기능을 제공합니다.

② 데이터 분석을 위해 필요한 주요 기능, 예를 들어 데이터셋 불러들이기, 데이터 전처리, 변수선택, 다양한 머신러닝 모델, 시각화 등 분석을 위한 툴박스입니다.

③ 툴박스에 있는 기능들을 마우스로 가져다가 분석을 진행할 수 있는 공간입니다.

④ 툴박스에서 선택한 기능에 대한 세부조정이 가능한 공간입니다.

2 데이터 세트



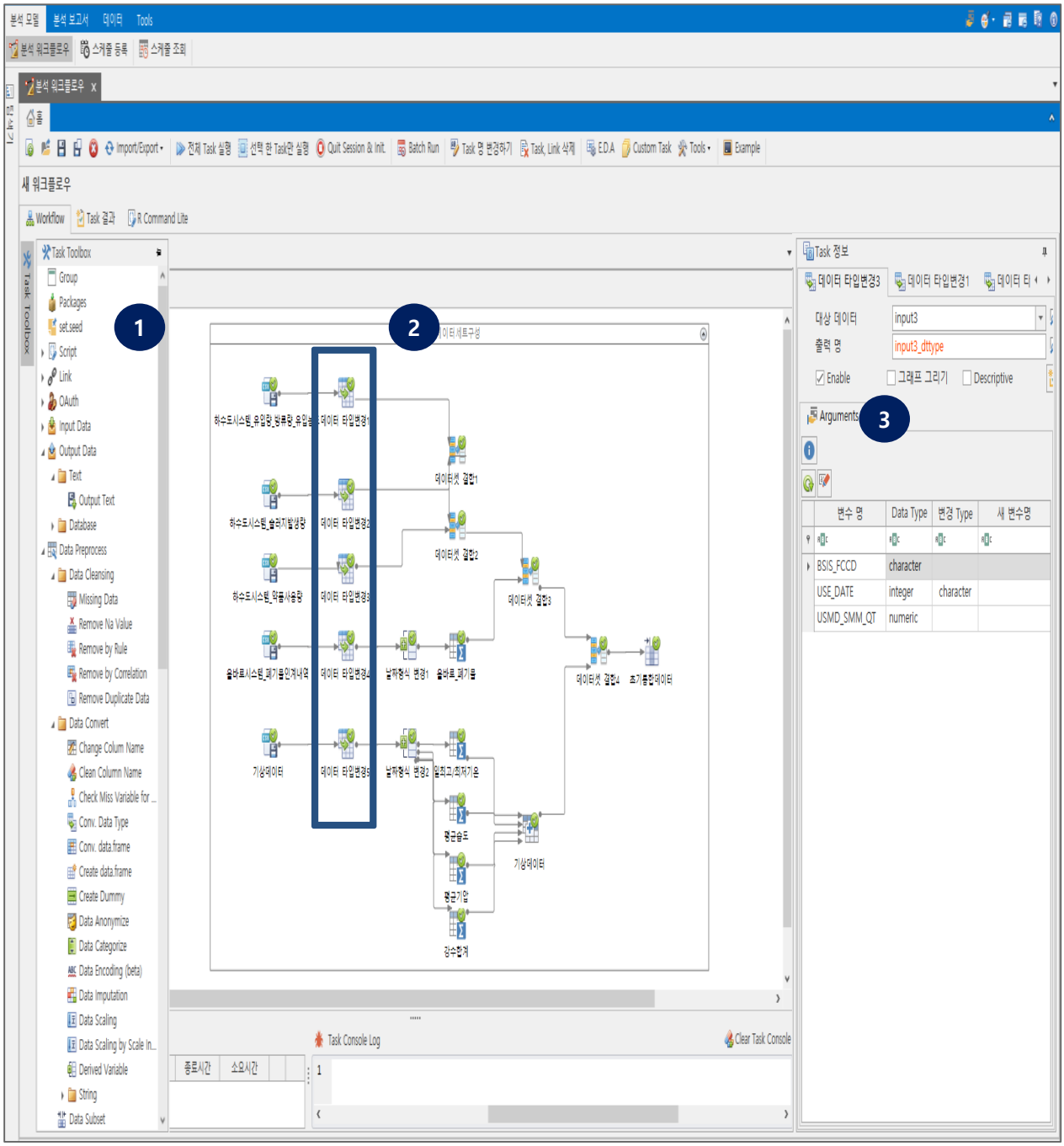
① 분석에 필요한 데이터셋을 불러들이기 위해 Task Toolbox에서 Input Data 폴더에서 이미지를 선택합니다.

② 선택한 이미지를 작업 공간에 놓습니다.

③ 데이터파일 경로, 출력명을 입력하고 데이터셋의 변수명, 변수타입을 확인합니다.

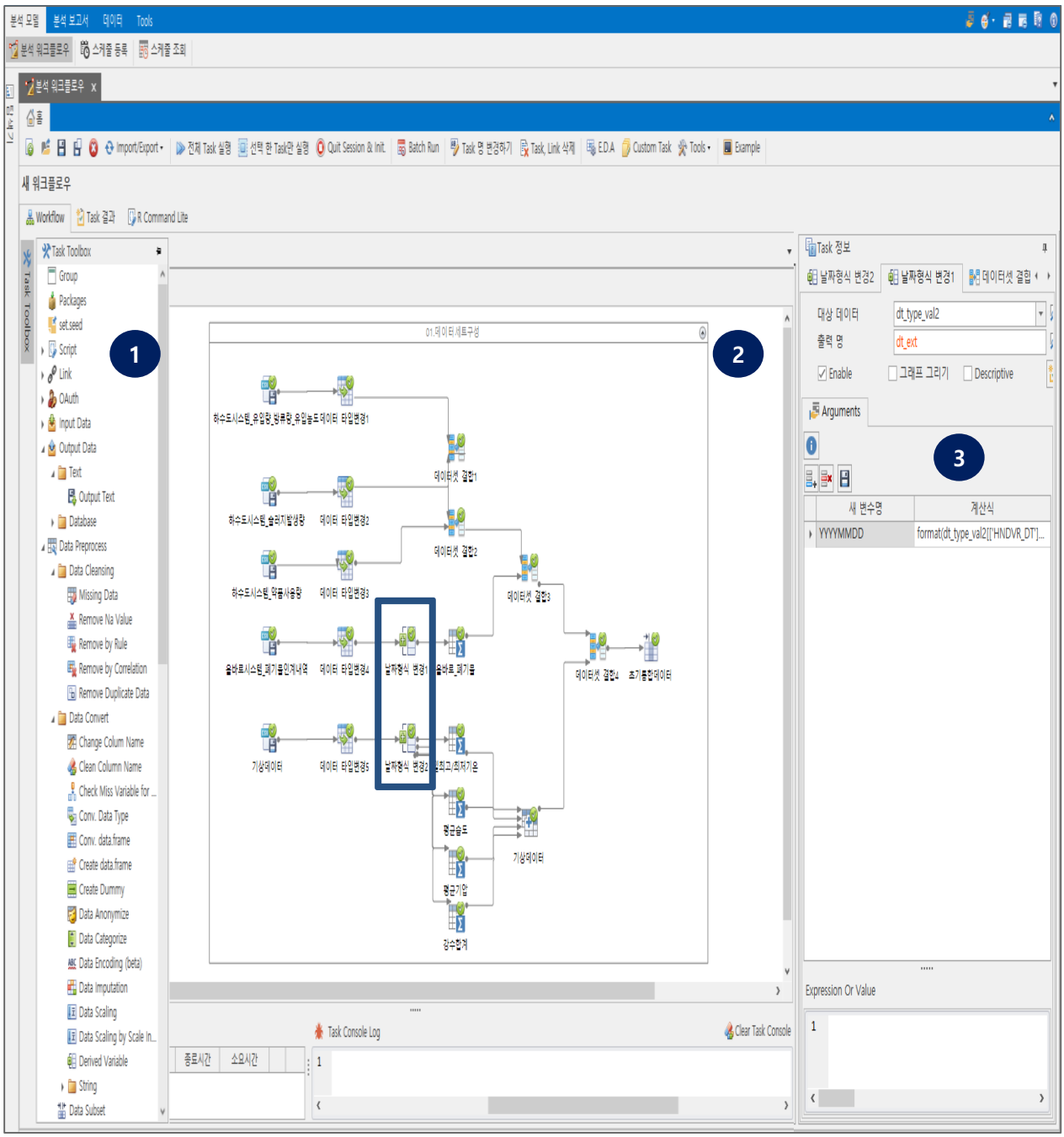
3 데이터 전처리

가. 데이터 타입 변경



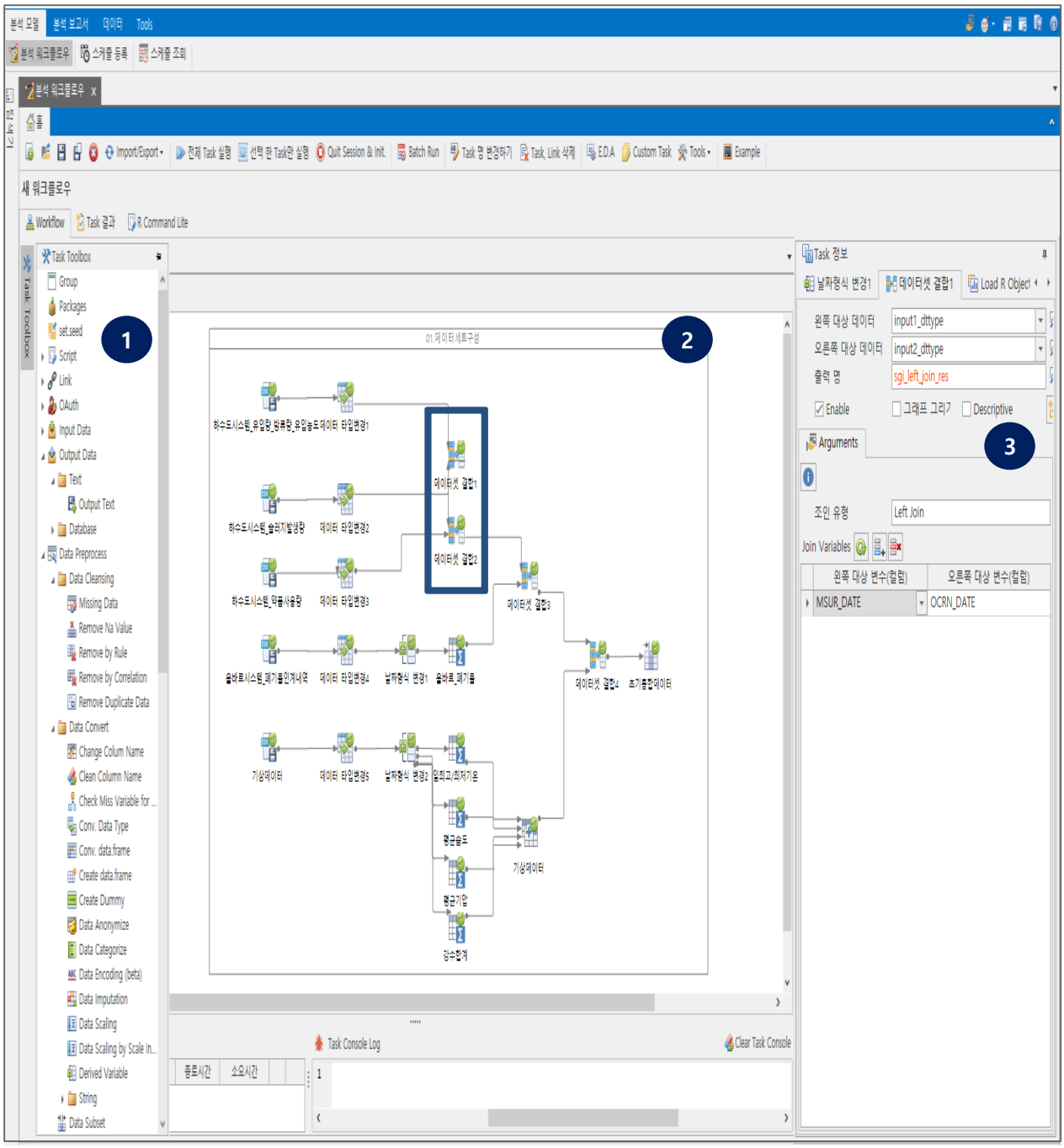
① 데이터 타입 변경을 위해 툴박스에서 Data Convert폴더의 Conv. Data Type 이미지를 선택합니다.
② 선택한 이미지를 작업공간에 놓습니다.
③ Task정보에서 각 변수들의 타입을 원하는 형태로 지정할 수 있습니다.

나. 파생변수 추가



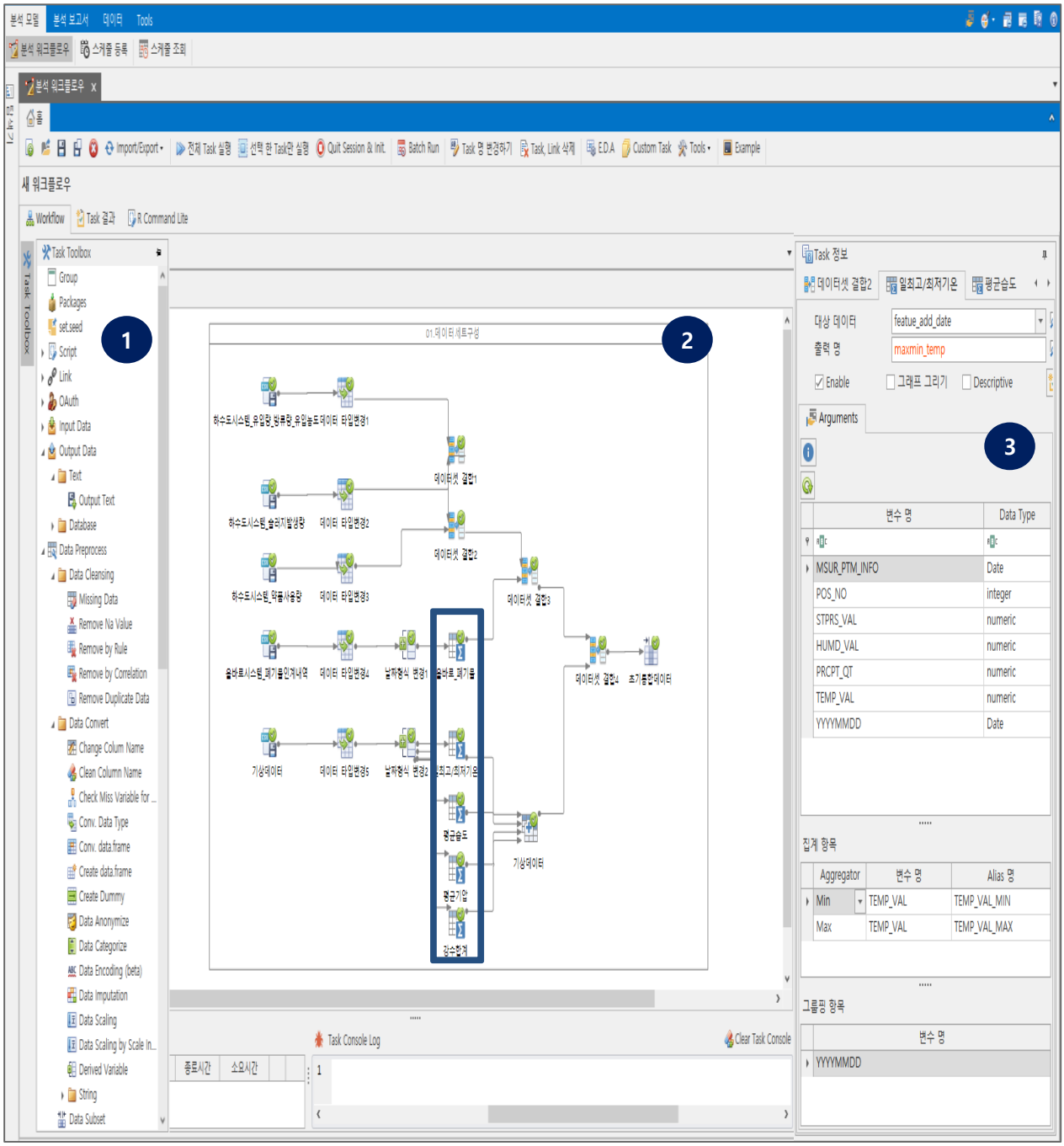
① 데이터 타입 변경을 위해 툴박스에서 Data Convert폴더의 Derived Variable 이미지를 선택합니다.
② 선택한 이미지를 작업공간에 놓습니다.
③ Task정보에서 원하는 형태의 파생변수를 추가할 수 있습니다.

라. 데이터셋 결합(Merge)



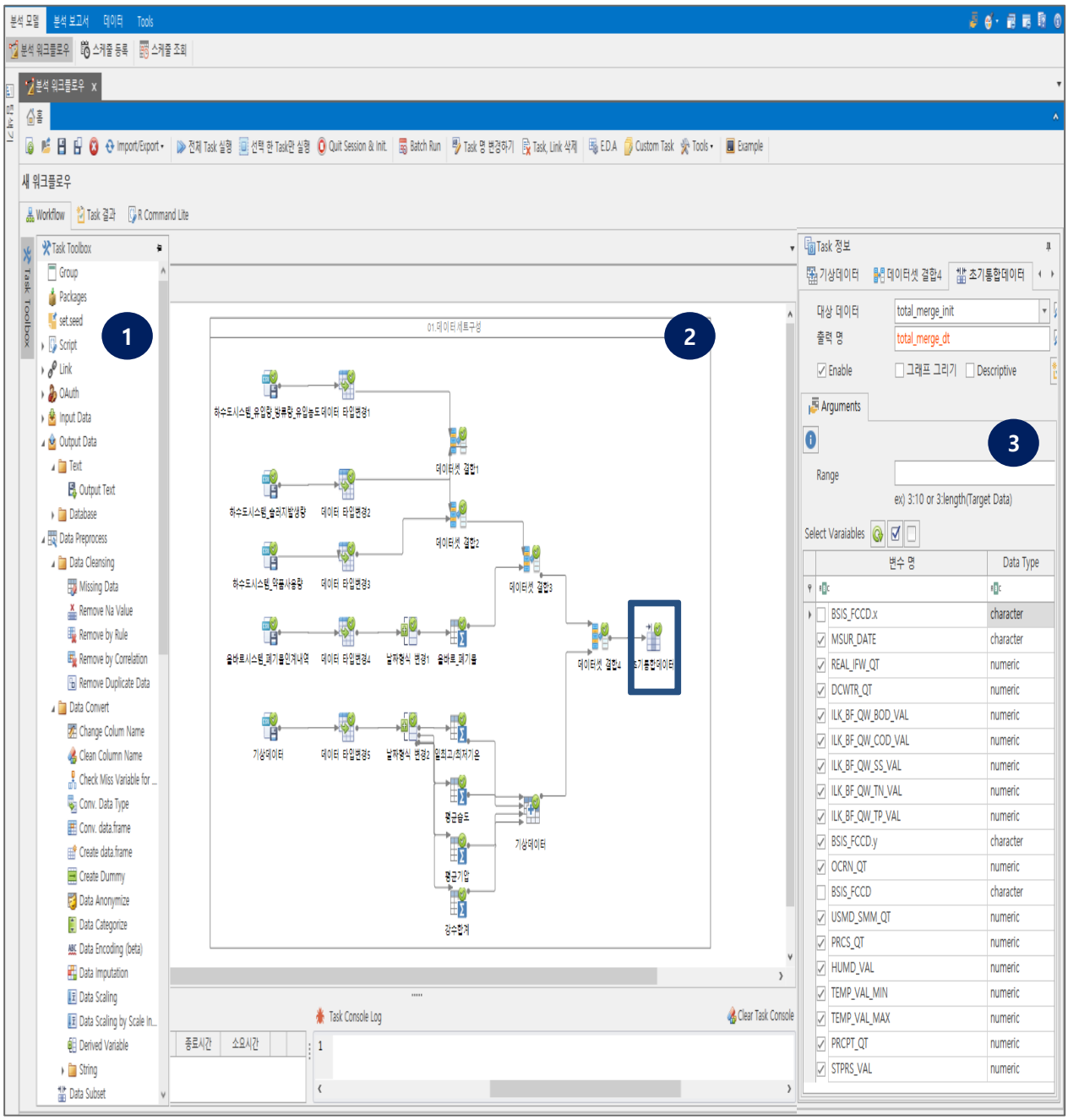
① 데이터 타입 변경을 위해 툴박스에서 Data Convert폴더의 Data Join 이미지를 선택합니다.
② 선택한 이미지를 작업공간에 놓습니다.
③ Task정보에서 Key변수를 지정하고 결합형태(Left, Right, Inner, Outer)를 선택할 수 있습니다.

마. 데이터셋 리샘플링(Aggregating / Resampling)



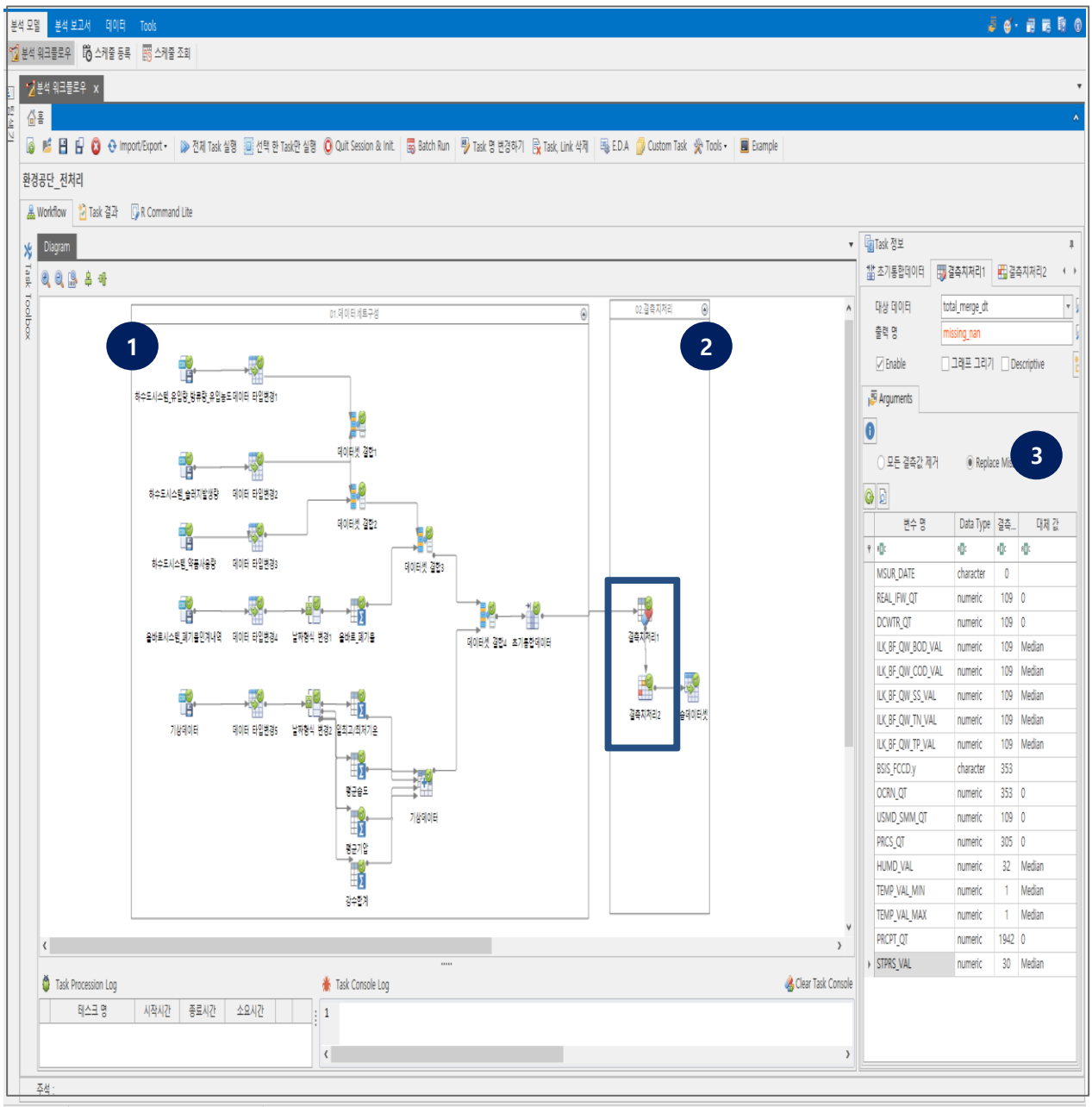
① 데이터 타입 변경을 위해 툴박스에서 Data Convert폴더의 Data Aggregate 이미지를 선택합니다.
② 선택한 이미지를 작업공간에 놓습니다.
③ Task정보에서 날짜 변수를 지정하고 필요한 집계번수들에 대해 최소값, 최대값, 합계, 평균, 갯수, 표준편차 등 선택 할 수 있습니다.

바. 데이터셋 통합구성



① 데이터 타입 변경을 위해 툴박스에서 Data Convert폴더의 Data Subset이미지를 선택합니다.
② 선택한 이미지를 작업공간에 놓습니다.
③ Task정보에서 결합된 데이터프레임에서 필요한 변수들을 선택하여 초기 통합데이터셋을 생성할 수 있습니다.

사. 데이터셋 결측치 처리



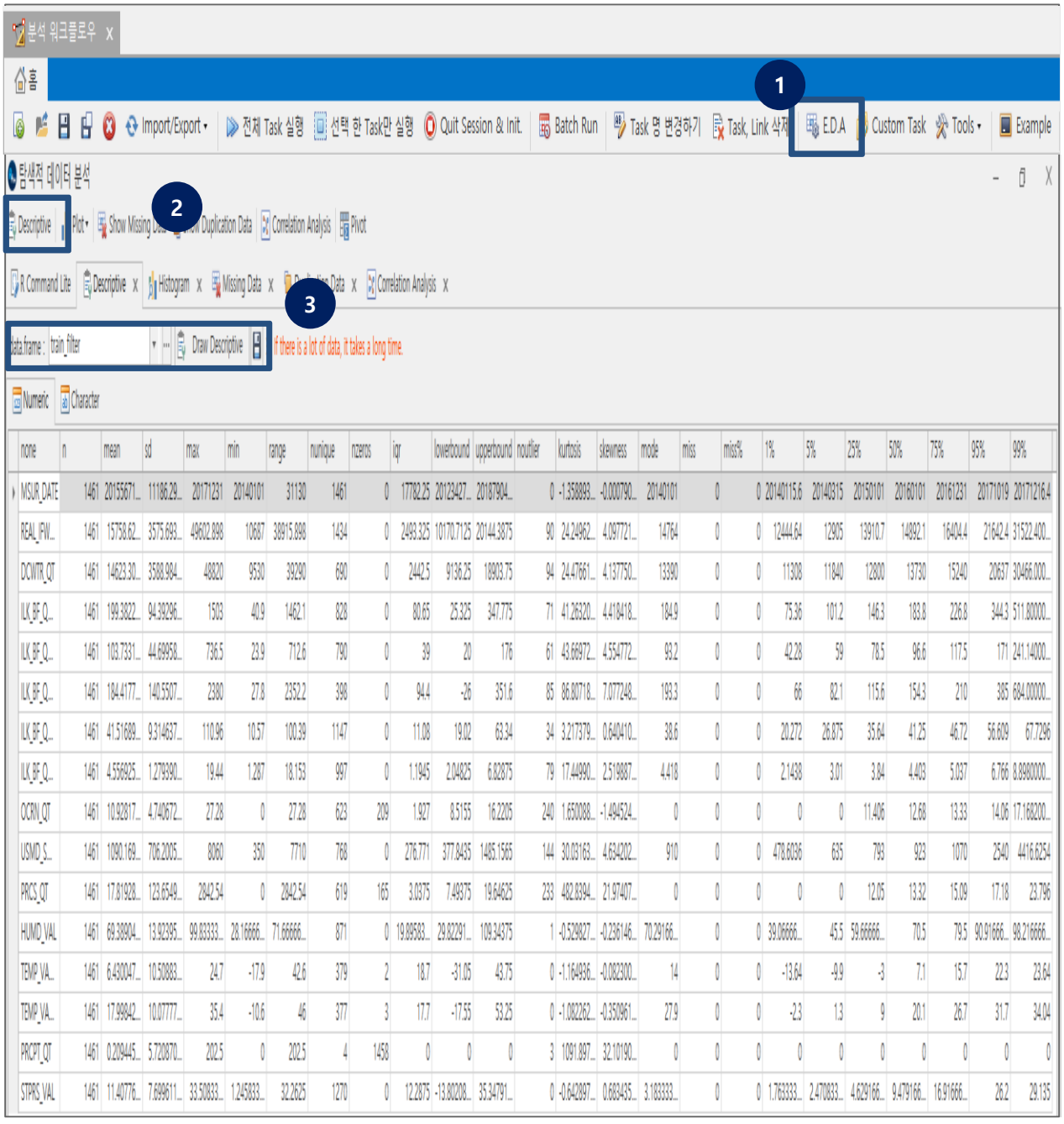
① 데이터 타입 변경을 위해 툴박스에서 Data Preprocess폴더의 Missing Data, Data Convert폴더의 Data Imputation 이미지를 선택합니다.

② 선택한 이미지를 작업공간에 놓습니다.

③ Missing Data이미지의 경우 평균, 중앙값, 최대값, 특정값을 채울 수 있으며, Data Imputation이미지의 경우 특정 함수를 지정하여 값을 채울 수 있다.

4 탐색적 데이터분석(Exploratory Data Analysis)

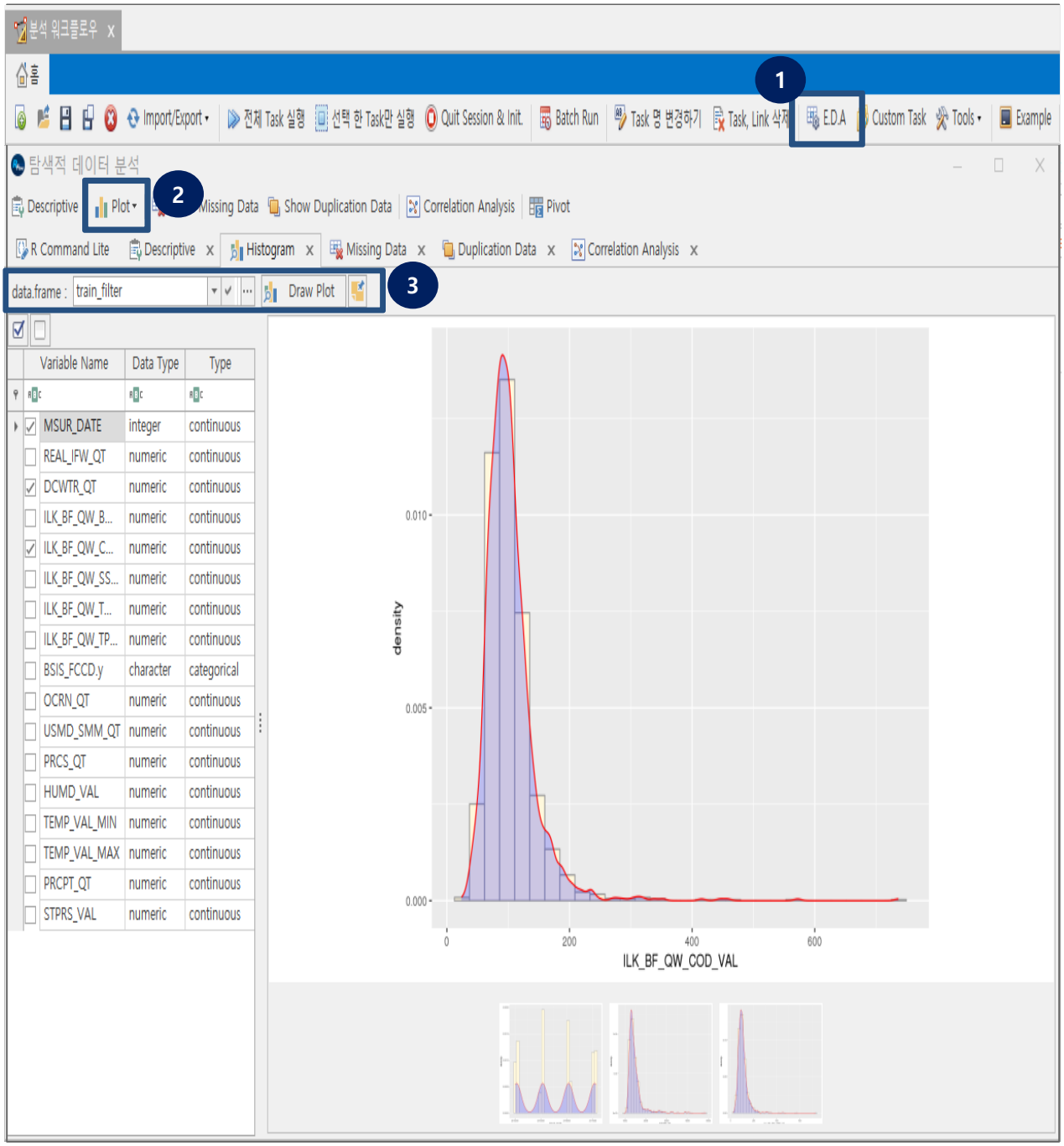
가. 데이터 통계량 분석



none	n	mean	sd	max	min	range	nunique	nzeros	iqr	lowerbound	upperbound	noutlier	kurtosis	skewness	mode	miss	miss%	1%	5%	25%	50%	75%	95%	99%
MSUR_DATE	1461	20155671...	11186.29...	20171231	20140101	31130	1461	0	17782.25	20123427...	20187904...	0	-1.358883...	-0.000790...	20140101	0	0	20140115.6	20140315	20150101	20160101	20161231	20171019	20171216.4
REAL_IPW...	1461	15758.62...	3575.693...	49602.888	10687	38915.888	1434	0	2493.325	10170.7125	20144.3875	90	24.24962...	4.097721...	14764	0	0	12444.64	12805	13910.7	14892.1	16404.4	21642.4	31522.400...
DCWTR_QT	1461	14623.30...	3588.984...	48820	9530	39290	690	0	2442.5	9136.25	18893.75	94	24.47661...	4.137750...	13390	0	0	11308	11840	12800	13730	15240	20637	30466.000...
ILK_BF_Q...	1461	199.3822...	94.39296...	1503	40.9	1462.1	828	0	80.65	25.325	347.775	71	41.26320...	4.418418...	184.9	0	0	75.36	101.2	146.3	183.8	226.8	344.3	511.80000...
ILK_BF_Q...	1461	103.7331...	44.68958...	736.5	23.9	712.6	790	0	39	20	176	61	43.66972...	4.554772...	93.2	0	0	42.28	59	78.5	96.6	117.5	171	241.14000...
ILK_BF_Q...	1461	184.4177...	140.5307...	2380	27.8	2352.2	398	0	94.4	-26	351.6	85	86.80718...	7.077248...	193.3	0	0	66	82.1	115.6	154.3	210	385	684.00000...
ILK_BF_Q...	1461	41.51689...	9.314637...	110.96	10.57	100.39	1147	0	11.08	19.02	63.34	34	3.217379...	0.640410...	38.6	0	0	20.272	26.875	35.64	41.25	46.72	56.609	67.7296
ILK_BF_Q...	1461	4.556925...	1.279390...	19.44	1.287	18.153	997	0	1.1945	2.04825	6.82875	79	17.44990...	2.519887...	4.418	0	0	2.1458	3.01	3.84	4.403	5.037	6.766	8.8980000...
OCRN_QT	1461	10.92817...	4.740672...	27.28	0	27.28	623	209	1.927	8.5155	16.2205	240	1.650088...	-1.494524...	0	0	0	0	0	11.406	12.68	13.33	14.06	17.168200...
USMD_S...	1461	1080.169...	706.2005...	8060	350	7710	768	0	276.771	377.8455	1485.1565	144	30.03163...	4.634202...	910	0	0	478.6036	635	793	923	1070	2540	4416.8254
PRCS_QT	1461	17.81928...	123.6549...	2842.54	0	2842.54	619	165	3.0375	7.49375	19.64625	233	482.8394...	21.97407...	0	0	0	0	0	12.05	13.32	15.09	17.18	23.796
HUMD_VAL	1461	69.38904...	13.90395...	99.83333...	28.16666...	71.66666...	871	0	19.88583...	29.82291...	109.34375	1	-0.520827...	-0.236146...	70.28166...	0	0	39.06666...	45.5	59.66666...	70.5	79.5	90.91666...	98.21666...
TEMP_VA...	1461	6.430047...	10.50883...	24.7	-17.9	42.6	379	2	18.7	-31.05	43.75	0	-1.164936...	-0.082300...	14	0	0	-13.64	-9.9	-3	7.1	15.7	22.3	23.64
TEMP_VA...	1461	17.99842...	10.07777...	35.4	-10.6	46	377	3	17.7	-17.55	53.25	0	-1.082262...	-0.350961...	27.9	0	0	-2.3	1.3	9	20.1	26.7	31.7	34.04
PRCPT_QT	1461	0.209445...	5.720870...	202.5	0	202.5	4	1458	0	0	0	3	1091.897...	32.10190...	0	0	0	0	0	0	0	0	0	0
STPRS_VAL	1461	11.40776...	7.699611...	33.50833...	1.245833...	32.2625	1270	0	12.2875	-13.80208...	35.34791...	0	-0.642897...	0.663435...	3.183333...	0	0	1.763333...	2.470833...	4.629166...	9.479166...	16.91666...	26.2	29.135

① 상위 화면 메뉴바에 E.D.A 아이콘을 선택합니다.
② 데이터 통계량을 확인하기 위해 Descriptive 아이콘을 선택합니다.
③ 데이터셋을 선택하고 Draw Descriptive 아이콘을 선택하면 위와 같이 데이터 통계량에 대한 표가 표현됩니다.

나. 그래프(Plot) 그리기

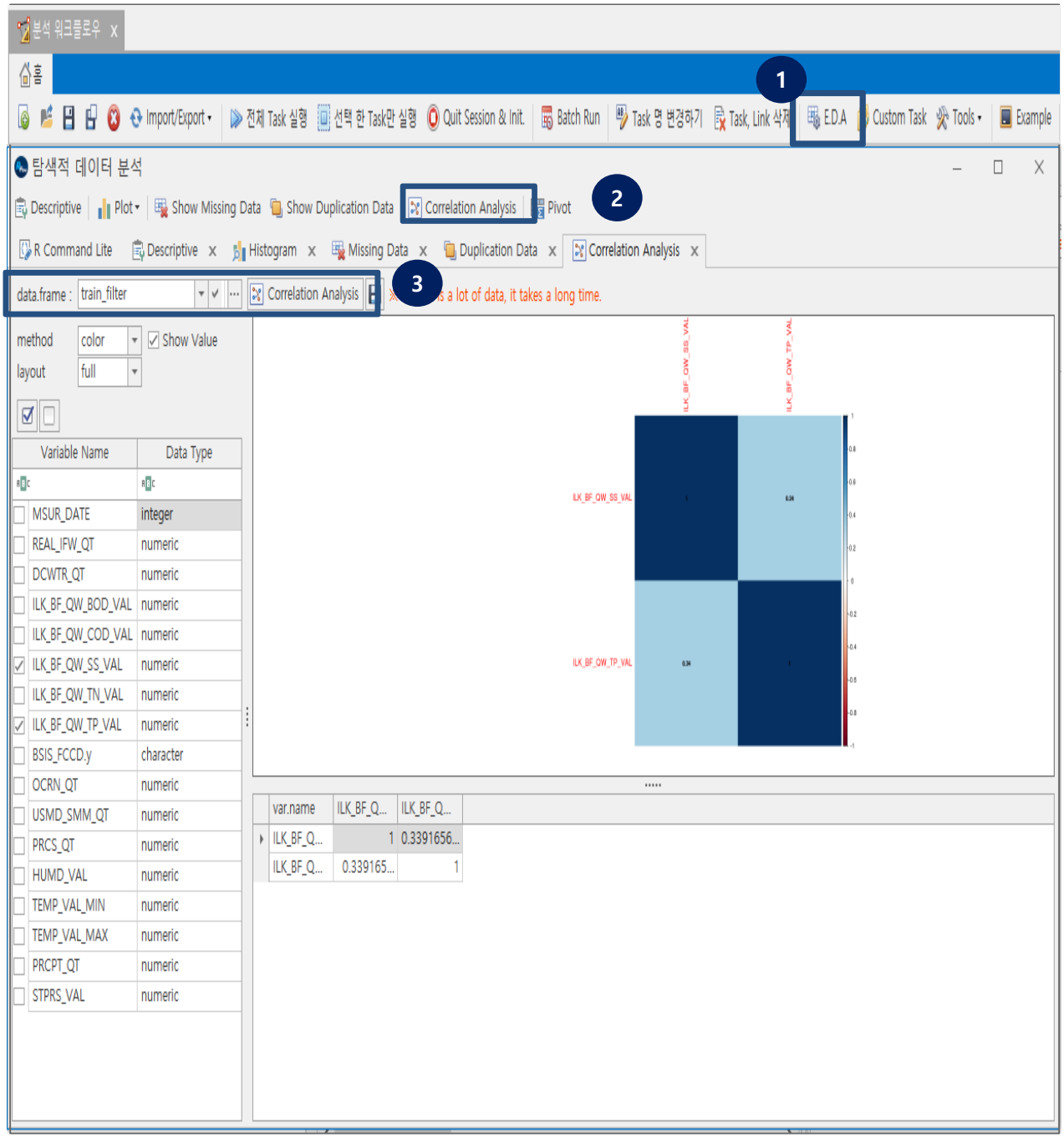


① 상위 화면 메뉴바에 E.D.A 아이콘을 선택합니다.

② 그래프를 그리기 위해 Plot 아이콘을 선택하면 Histogram, Boxplot, Scatter Plot을 지정할 수 있습니다.

③ 데이터셋과 변수들을 선택하고 Draw Plot 아이콘을 선택하면 위와 같이 그래프가 표현됩니다.

다. 데이터 상관계수 분석



① 상위 화면 메뉴바에 E.D.A 아이콘을 선택합니다.
② 상관계수 분석을 위해 Correlation Analysis 아이콘을 선택합니다.
③ 데이터셋과 변수들을 선택하고 Correlation Analysis 아이콘을 선택하면 위와 같이 데이터 상관계수에 대한 표와 그림이 표현됩니다.