I. Brightics 소개 및 사용법 실습(오픈클래스)

Studio

글씨가 잘 보이시나요? ← 진행하겠습니다.

강사 소개

에듀테이너 (Robert Yong Park, Ph.D.)

- 現. Principal Consultant/데이터분석팀/SDS (2023~)
- 前. Principal Engineer/Analytics팀/연구소(2014~)
- 前. 수석컨설턴트/SCM컨설팅,전략마케팅&SL사업부(2005~)
- 한국국방연구원, GaTech, SNU

* 사내블로그, LinkedIn 참조

Brightics Studio Open Class

구분	강의 내용	시간

세션 1	• 1교시 (10~) : Brightics 사용하기
	  • Brightics 소개
	  • 모델링기능
	  • 차트분석기능
		50 min.

	Break (10 min)	

세션 2	• 2교시(11~) : Brightics로 분석하기
	  • 기본시나리오
	  • 시나리오예제
		60 min.

QnA	※ Youtube BrighticsTV에서 실습 가능 https://youtu.be/0CpqKwMdtKM?si=5hk5PKAvM1fL7mMd	

참고. Brightics AI 관련 Sites

- Brightics 포털 - https://www.brightics.ai
  - ✓ 사용자 매뉴얼 : User Guide
  - ✓ 데이터분석에 대한 자율학습서 : Tutorial
  - ✓ 도메인별 샘플모델 : Use Cases
  - ✓ 차트: chart option guide
  - ✓ 함수에 대한 설명 : Function Reference
  - ✓ 질의/응답 : Question & Answer mailto → brightics.cs@samsung.com

- Brightics TV (youtube)
  - ✓ Brightics 사용법 동영상 교육자료
  - ✓ Brightics Instruction in English

***café.naver.com/brighticsstudio (개인)**

insight to !nspiration

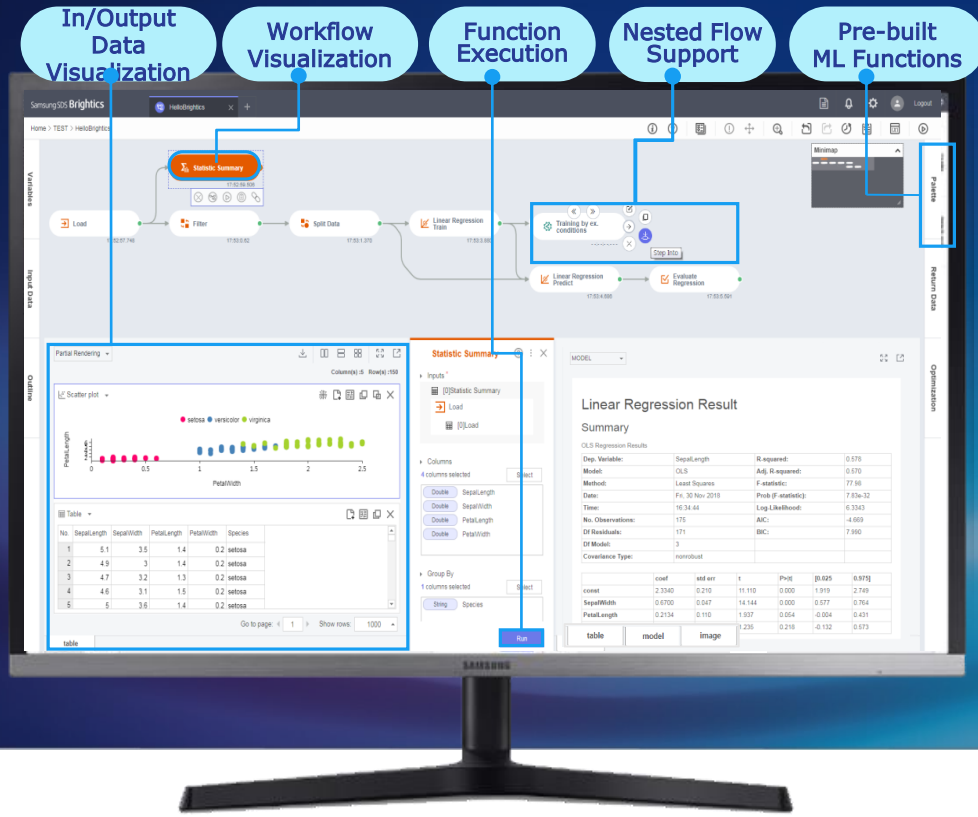
Samsung SDS

Brightics AI 소개

삼성SDS Brightics AI는
쉽고 빠른 데이터처리 및 분석모델링과 함께 경제적인 동시에
안정적인 분석모델관리환경을 제공하는 지능형 분석플랫폼 기반
기업데이터 분석서비스입니다.

삼성SDS SAMSUNG

쉽고, 빠르고, 똑똑한 AI 통합분석플랫폼



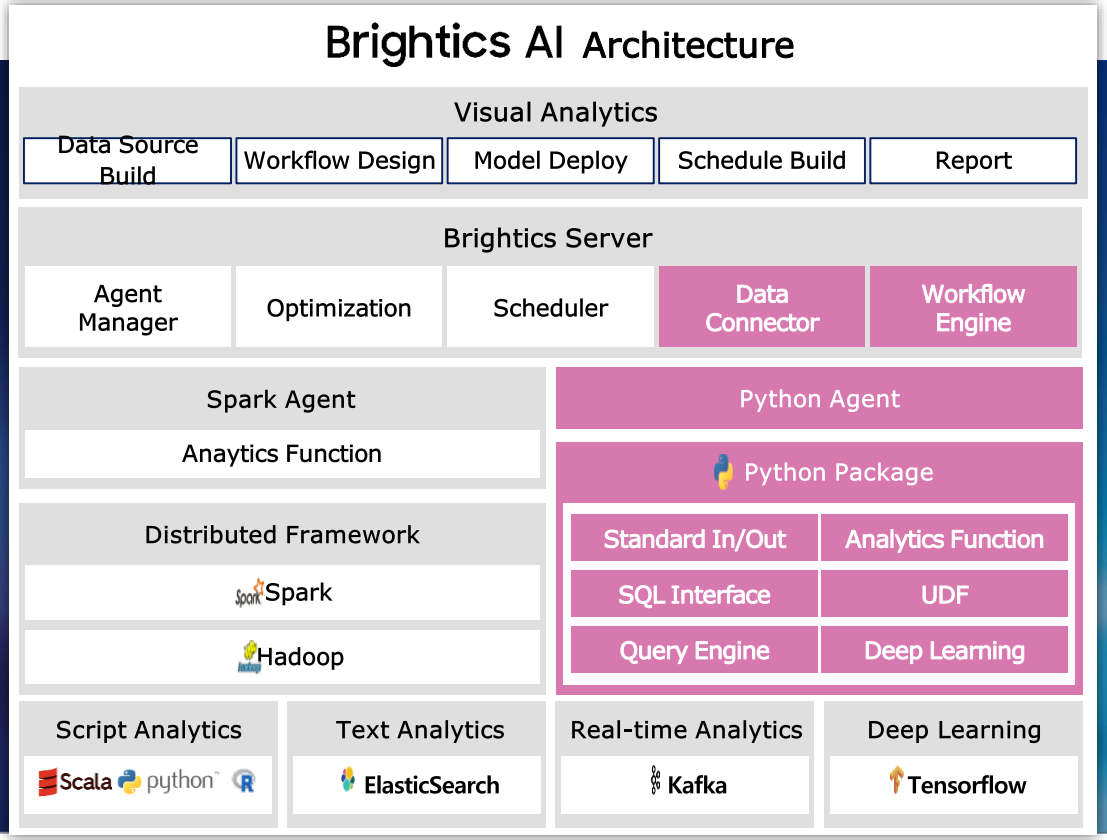
SAMSUNG SDS
Brightics AI

데이터 로딩부터 모델링, 분석 결과 리포트 배포까지 가능한
사용자 친화적 분석환경을 제공합니다.

Pre-built ML 함수 제공 및 Auto ML을 활용한 변수/알고리즘 자동추천으로 **모델링 시간이 획기적으로 단축됩니다.**

R, Python, SQL 등 분석언어 뿐 아니라 상용 BI, ETL 툴 연계로
확장성 높은 관리환경을 제공합니다.

고객 니즈에 맞춘 플랫폼 Edition



SAMSUNG SDS
Brightics AI

Spark 기반 특화 분산처리 기술로 빅 데이터 처리속도를 차별화한

Brightics Enterprise (ML)

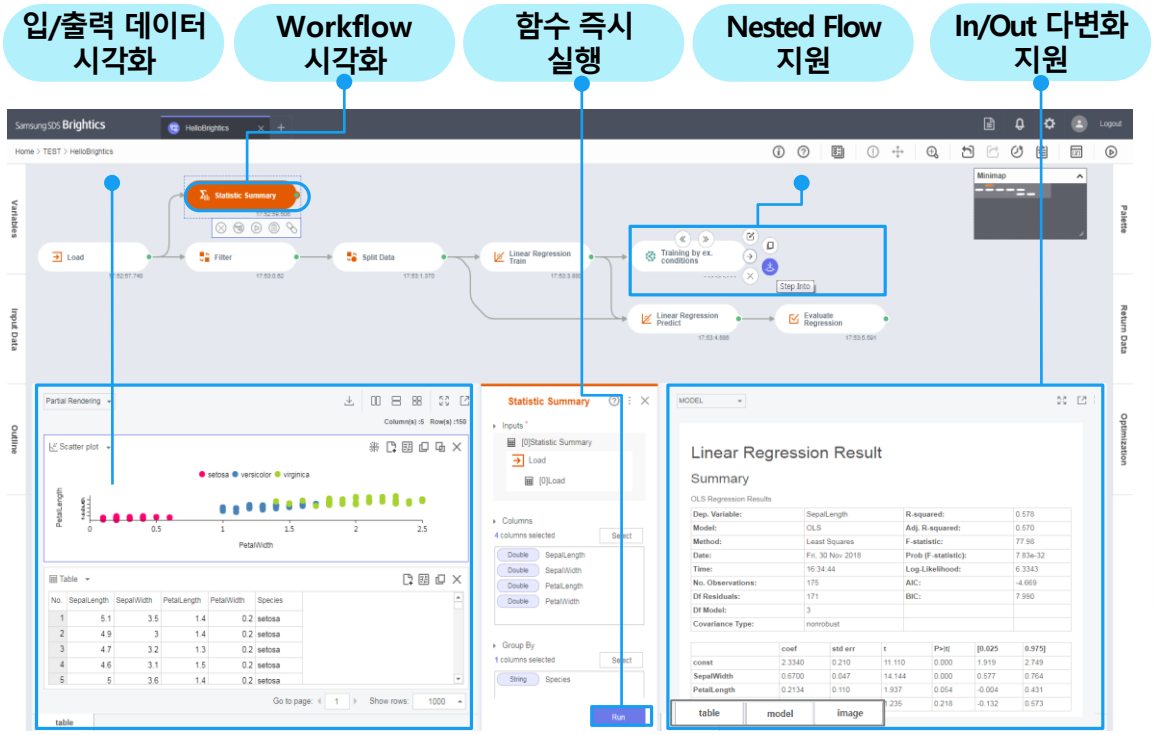
개인PC에서도 사용 가능한 오픈 소스 버전

Brightics Studio

Brightics AI 주요 특징

1 시각화된 통합 분석 환경

데이터와 분석workflow를 모두 한 화면에서 처리



주요특징

- **Workflow 시각화**
  분석의 흐름을 한눈에 파악할 수 있도록
  카테고리별 함수 분류,
  함수 수행 여부에 관한 상태 등 제공
- **Nested Flow(Subflow) 지원**
  subflow, If-else, Loop 기능 지원을 통한
  시각화된 분석방법론 강화
- **함수 즉시 실행**
  변수 값 변경 후 즉시 실행하여
  테스트 편의성 향상
- **In/Out 데이터 다변화**
  데이터를 분석 값에 따라 테이블, 모델, 이미지 등
  다양한 UI 레이아웃으로 제공

Brightics AI 주요 특징

2 Pre-built ML 함수

집계, 예측, 처방형 분석모델링을 Drag-and-drop 방식으로 지원

Brightics v1.0 | v2.0 | v3.0

Descriptive Analytics (집계)

Manipulation
Outlier Removal, Sort, Filter, Replace Missing Number, etc.

Transform
Type Cast, Join, Bind Column, Pivot, PCA, Split Data, etc.

Statistics
Association Rule, Correlation, Frequency, T-Test, etc.

Extraction
Moving Average, EWMA, Normalization, etc.

Clustering
Gaussian Mixture, K-means, etc.

Predictive Analytics (예측)

Regression
Linear, Isotonic, GLM, Stepwise, etc.

Time Series
Auto ARIMA, Holt-Winters, etc.

Evaluation
Evaluate Regression, Ranking, Binary Classification, etc.

Text Analytics
Elastic Indexing, Elastic Search, Elastic Query, etc.

Recommendation
ALS Recommend

Classification
Decision Tree, K-NN, SVM, Naïve Bayes, Logistic, etc.

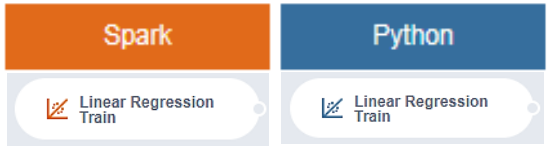
Prescriptive Analytics (처방)

Autonomous
EDA, Cleansing, Regression, Time Series, Classification, etc.

Optimization
Local/Global Optimization, Parameter Studies, etc.

Deep Learning
Core, Convolutional, Pooling, Recurrent Layer, etc.

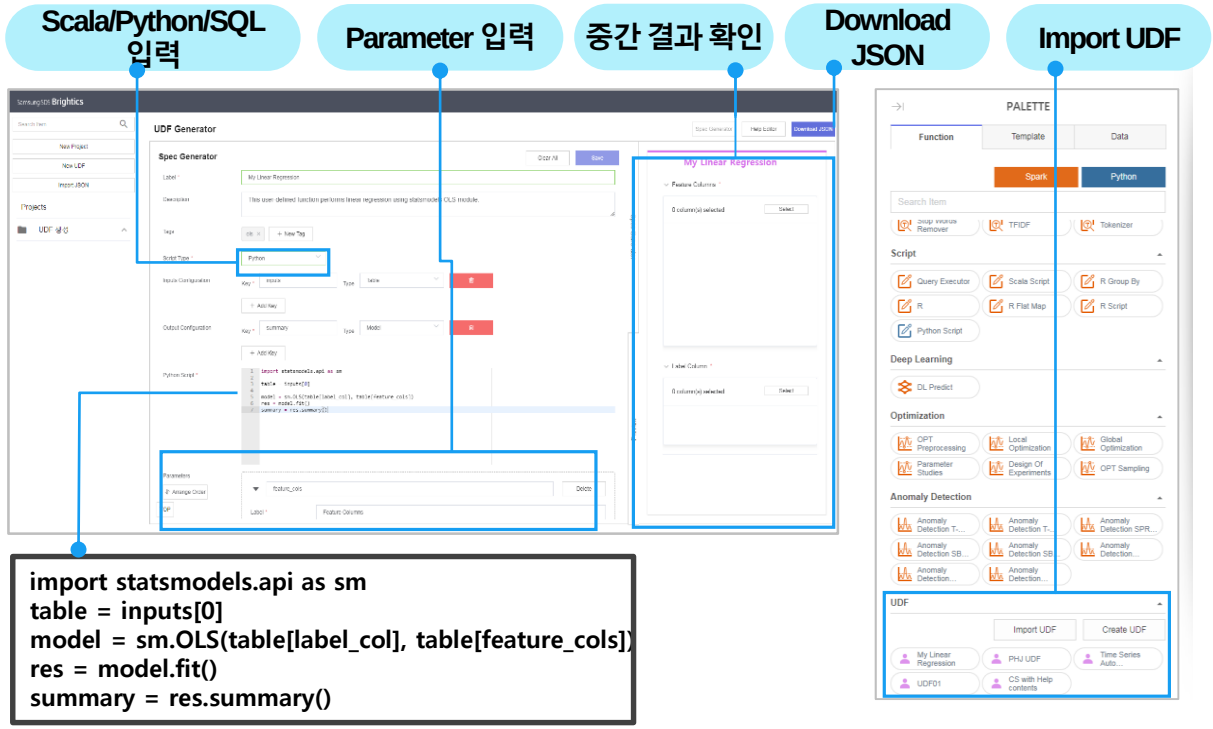
Spark기반의 대용량 함수(총 237종)와
Python기반의 중소용량(총 220종) 함수 지원



Brightics AI 주요 특징

3 사용자 정의 ML 함수 : User Defined Function 기능

분석에 필요한 함수를 사용자가 직접 만들어 Brightics에 탑재 가능



주요특징

- **다양한 Script language 지원**
  Scala/Python/SQL 중에 사용자가 편리한 언어를 선택하여 함수 개발 가능

- **Parameter 컨트롤 지원**
  Input, Array Input, Expression, Column Selector, Checkbox, Boolean Radio Button, Radio Button, Dropdown List, File Selector 등 다양한 타입의 파라미터 생성 가능 (총 9종)

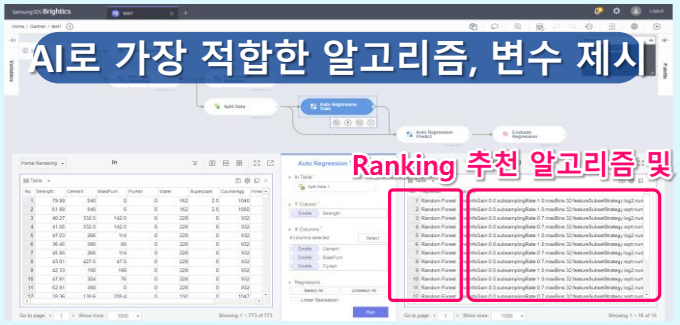
- **예측 결과 화면 제공**
  작업 도중에 틈틈이 현재까지 작업한 결과물 UI를 오른쪽 디스플레이 패널에서 바로 확인

4 Auto ML : 자동 변수/분석알고리즘 추천

데이터 특성에 맞는 최적알고리즘 자동추천으로 반복적인 모델링 시간 단축

Manual
Analytics

Autonomous
Analytics



예시 빌딩 전력 수요예측

주요특징

- **최적 알고리즘 자동추천**
  데이터 특성에 맞는 가장 적합한 알고리즘, 변수 추천
- **다양한 자동 분석함수**
  Data Cleansing, EDA, Regression, Classification, Time Series 등 각 영역별 자동 분석알고리즘 제공 (총 20종)
- **Optimization 기능 제공**
  Local / Global Optimization, Parameter Studies, Design of Experiments 등 최적화 알고리즘 6종 제공

Brightics AI 주요 특징

4 Script Modeling : R, Python, SQL, Scala

Script를 이용한 REPL[1] 분석이 가능한 데이터 분석환경 제공

주요특징

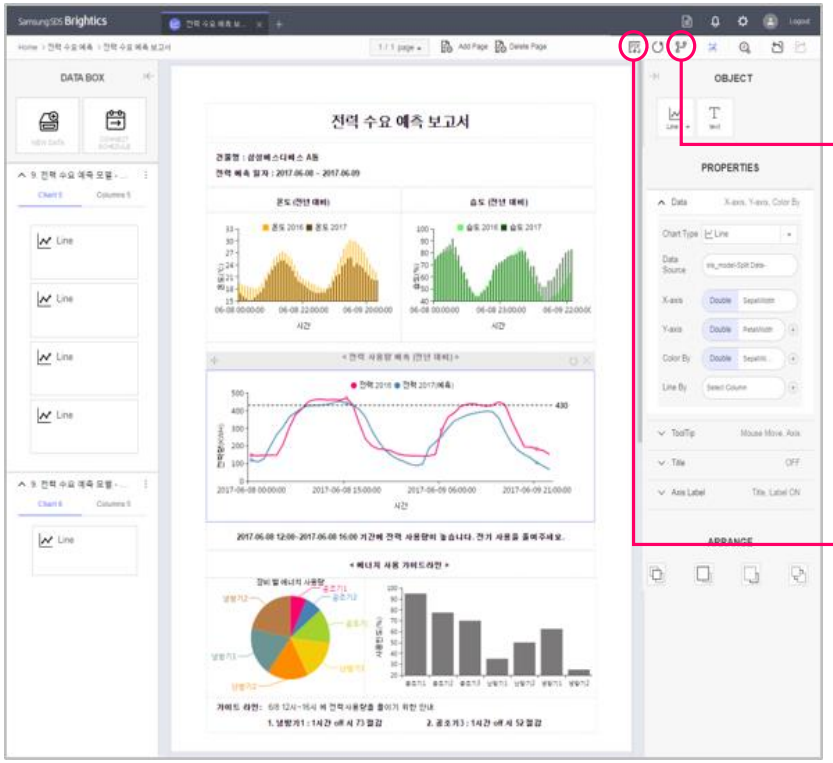
- **다양한 Script 지원**
  SQL, Scala, R, Python 등
  다양한 Script 사용환경 제공
- **Interactive Modeler**
  Trial & Error 를 기반으로 data 탐색이
  가능한 분석환경 제공
- **결과 시각화**
  다양한 타입의 Multi Charts 기능 제공
- **함수 별 예제 Script**
  Script 작성 시간 단축을 위해 미리
  제공되는 대용량 분석 함수

[1]REPL : Read-Eval-Print Loop

Brightics AI 주요 특징

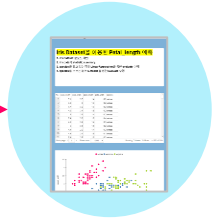
5 분석 리포트 자동생성 및 배포

간편한 리포트 편집기능 과 분석리포트 배포 (웹/모바일)



웹 (URL, HTML)

: E-mail, 블로그 등 으로 배포가능



PDF 출력

주요특징

- **편리한 리포트 편집**
  31종 차트를 통해 전체, 일부 데이터를 다양한 관점으로 표현하여 리포트로 제공
- **보고서 배포 (웹, 모바일)**
  URL, HTML Embedding 등
  3rd Party Application에서 Report 확인가능
- **주기적 Report 자동생성**
  Scheduler에 설정된 주기에 맞춰 자동생성 및 배포 진행
- **PDF 출력**
  작성한 보고서를 PDF파일로 export하여 출력


Thank you.

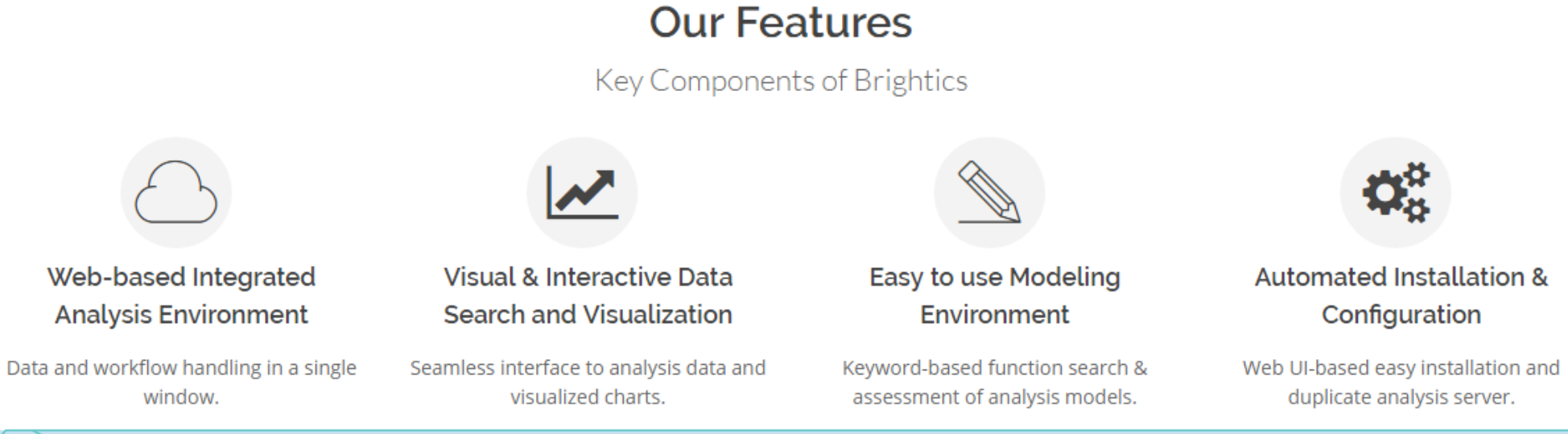
1. Brightics Visual Analytics 활용법

Brightics Studio 사용법 교육
1 모델링

https://www.youtube.com/watch?v=0CpqKwMdtKM&ab_channel=BrighticsTV

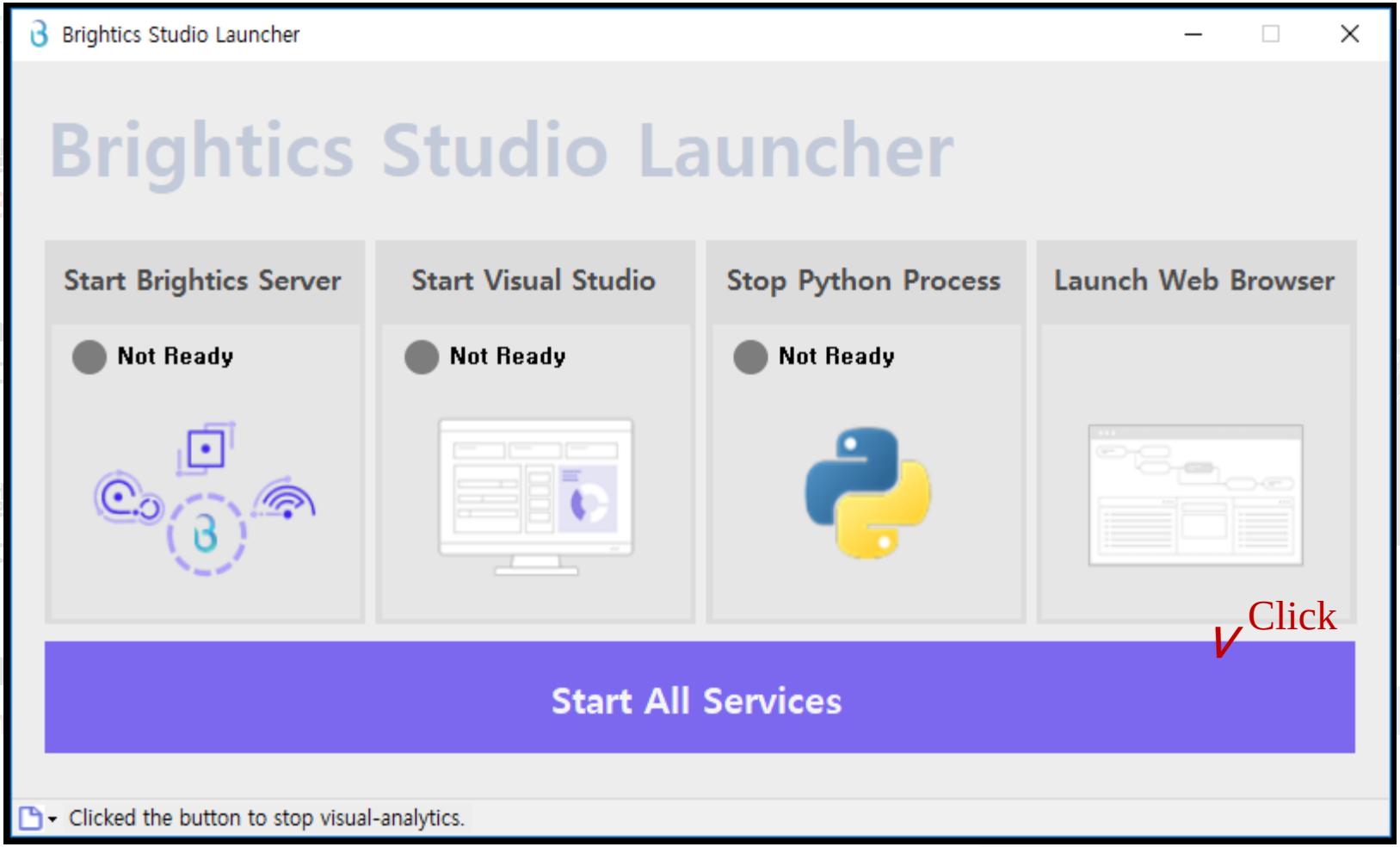
Brightics Studio

- 교육 대상: Brightics Studio를 처음 사용하는 사람
- 교육 목표
  1. Brightics Studio의 화면(메뉴) 구성에 대해서 이해
  2. 교육용 샘플 데이터 사례로 modeling하면서 Brightics Studio 사용을 연습
  3. Brightics Studio 리포트 작성법 이해



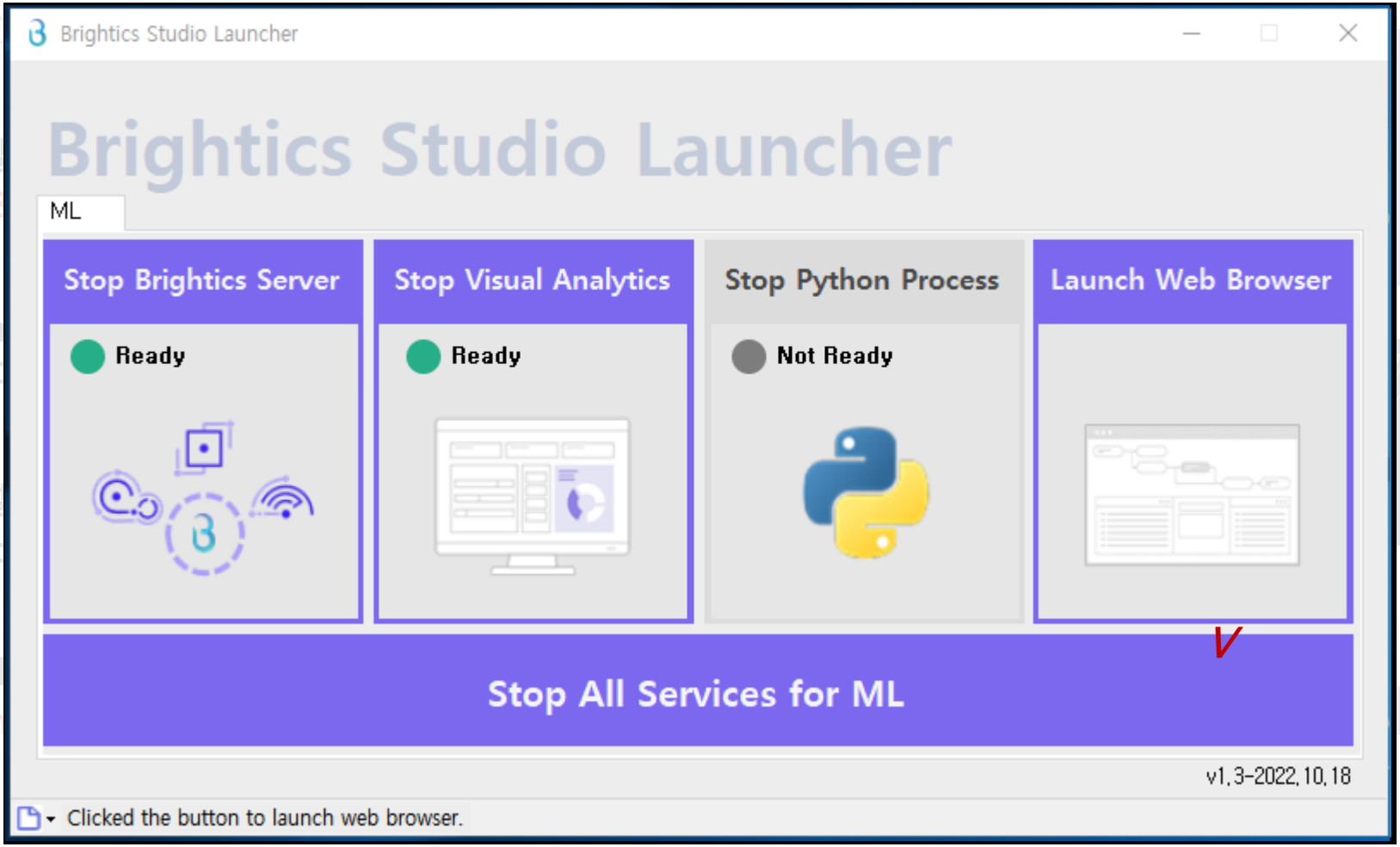
Brightics Studio Start

■ Start All Services for ML 을 클릭하여 메모리에서 실행



Brightics Studio Stop

■ Stop All Services for ML 을 클릭하여 메모리에서 STOP



※ Stop을 안 하면 메모리를 계속 차지함

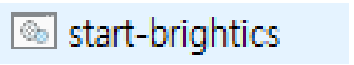
Brightics 시작하기

Brightics Studio의 화면구성과

기본 메뉴 아이콘에 대하여 간략히 알아봅니다

이후 샘플데이터로 실습을 진행합니다.

클릭

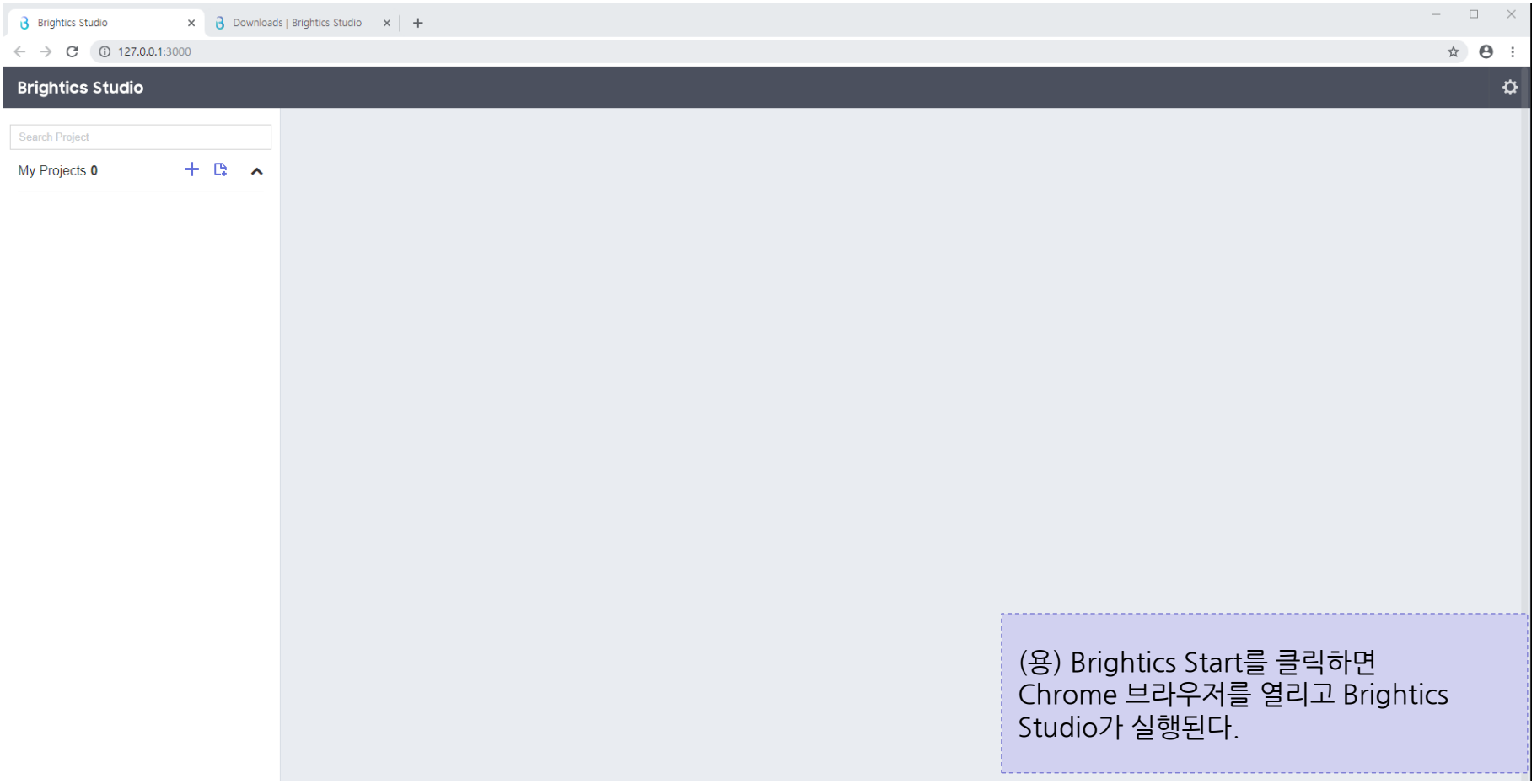


실행

화면 구성 (1)

URL : http://127.0.0.1:3000

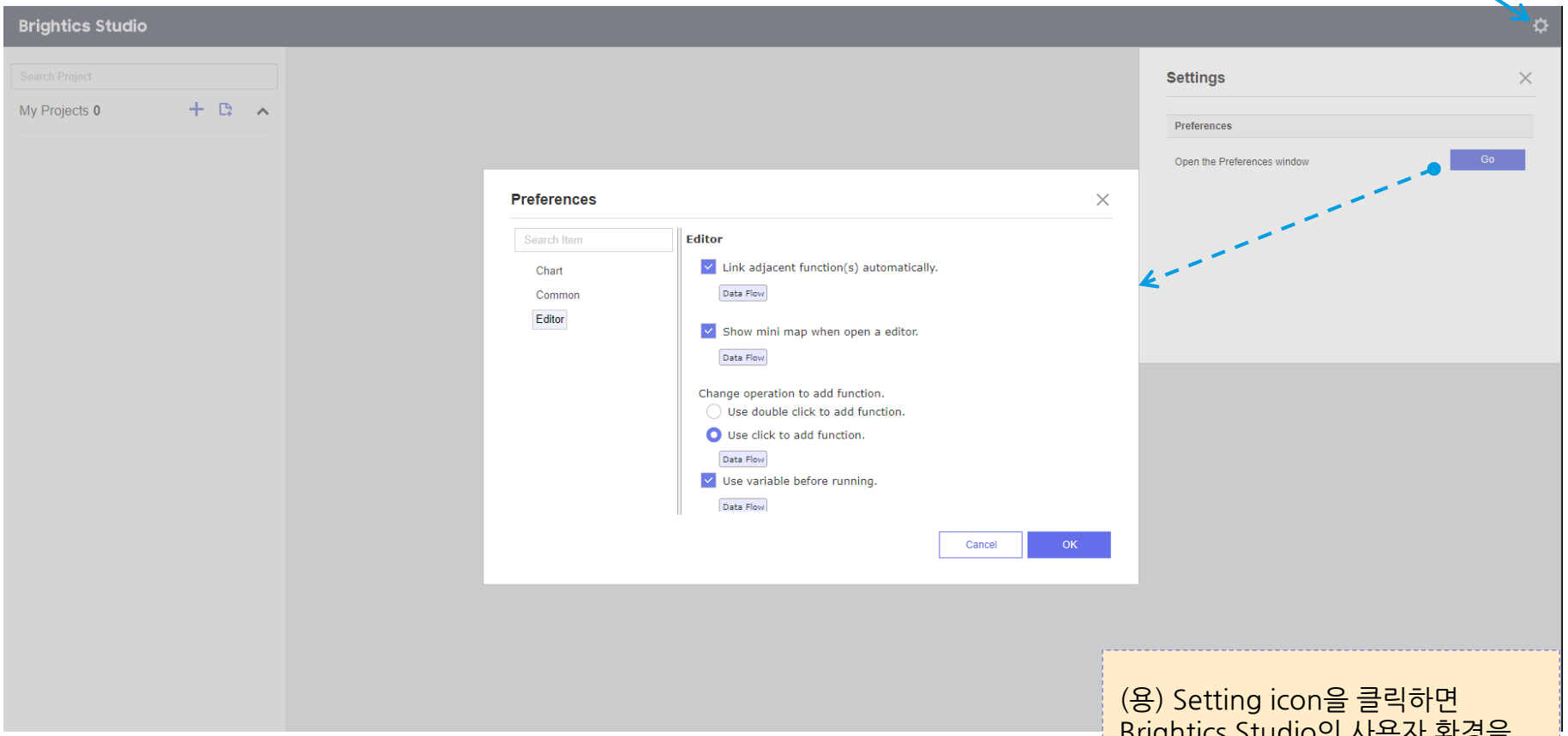
ID / PW : Local



(용) Brightics Start를 클릭하면 Chrome 브라우저를 열리고 Brightics Studio가 실행된다.

화면 구성 (1)

- **Preferences Setting.**

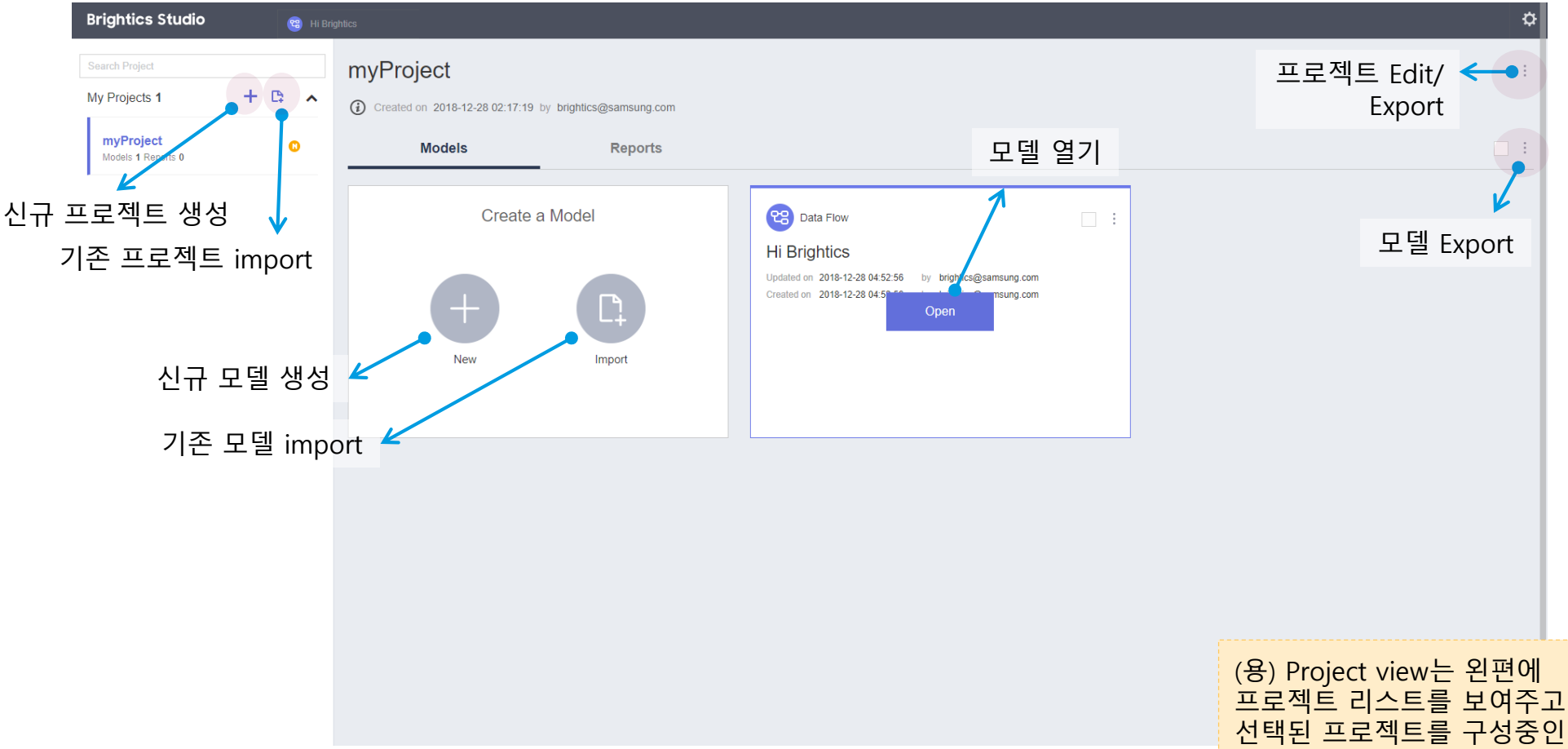


(용) Setting icon을 클릭하면 Brightics Studio의 사용자 환경을 설정할 수 있다.

화면 구성 (2) - Project view

프로젝트 관리 : Description 수정/삭제, 프로젝트 추가/삭제

모델 관리 : Description 수정/삭제, 모델 Export

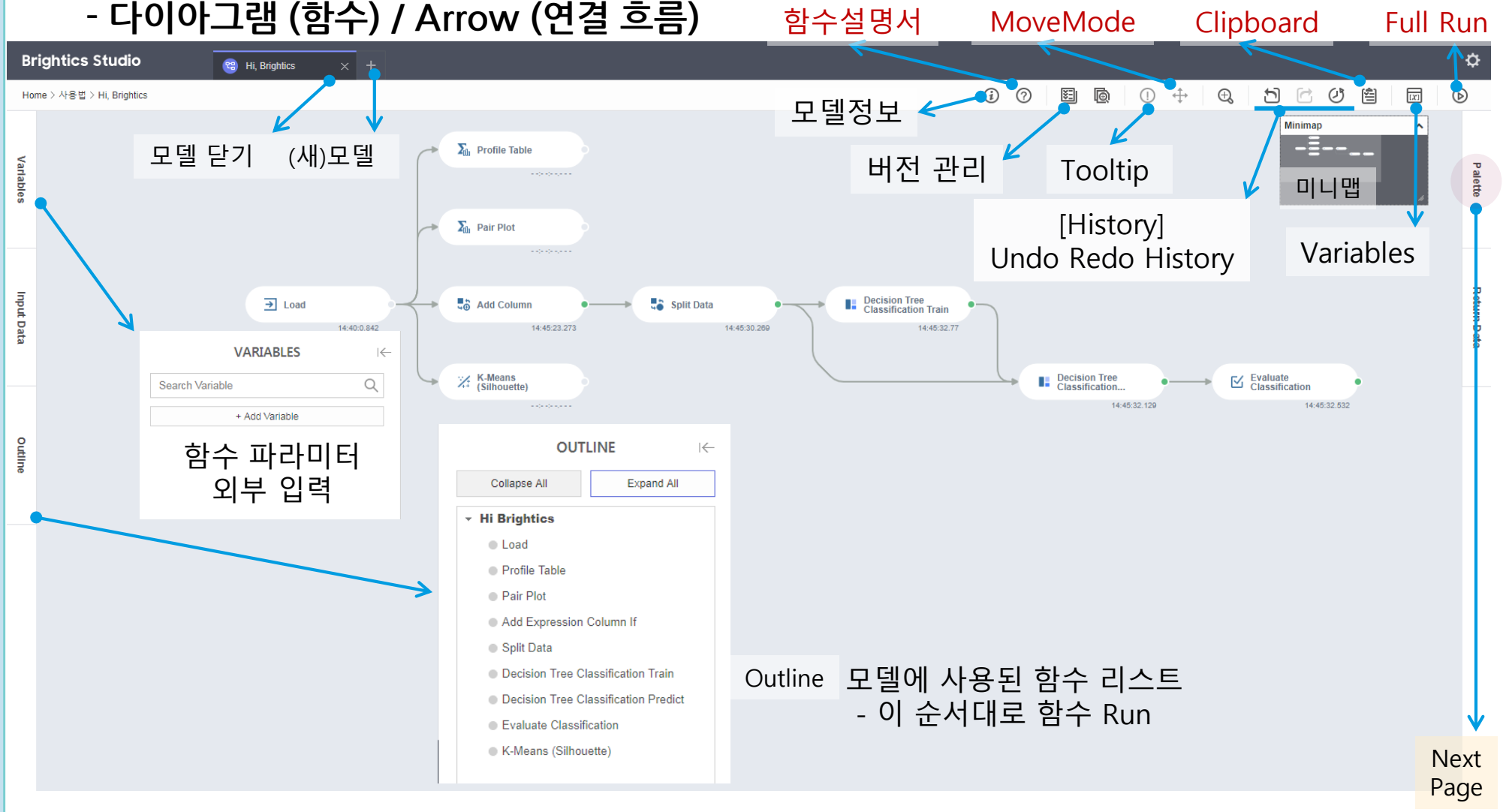


(용) Project view는 왼편에 프로젝트 리스트를 보여주고 선택된 프로젝트를 구성중인 모델/리포트를 보여준다.

화면 구성 (3) - Model View(1)

모델 내에 사용된 함수들이 Data Flow로 연결됨

- 다이아그램 (함수) / Arrow (연결 흐름)

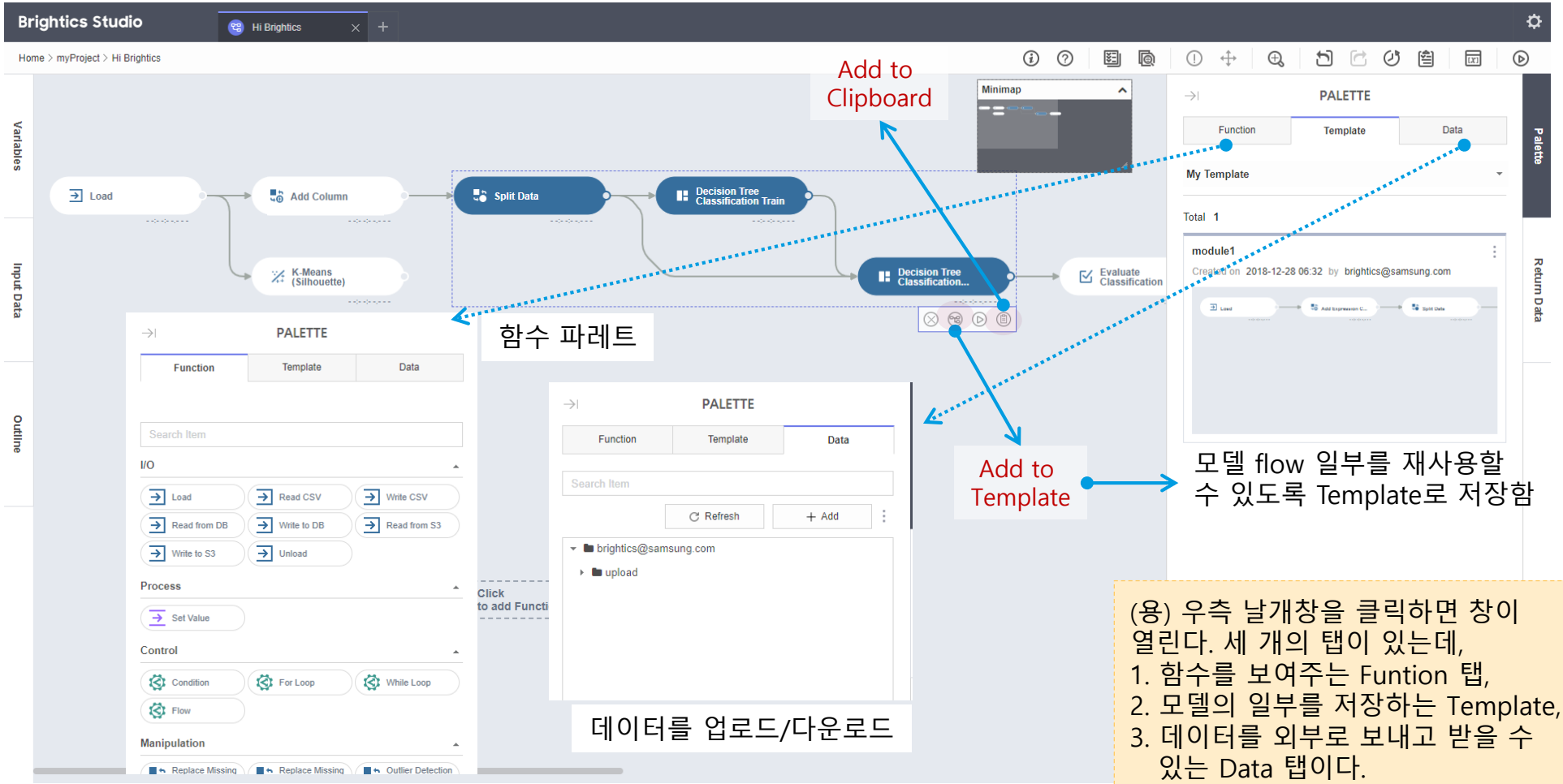


화면 구성 (3) - Model View(2), Diagram Editor

PALETTE (오른쪽 날개창)

- Function / Template / Data

Clipboard는 로그아웃시 삭제됨.
Template는 Name으로 자동 저장됨

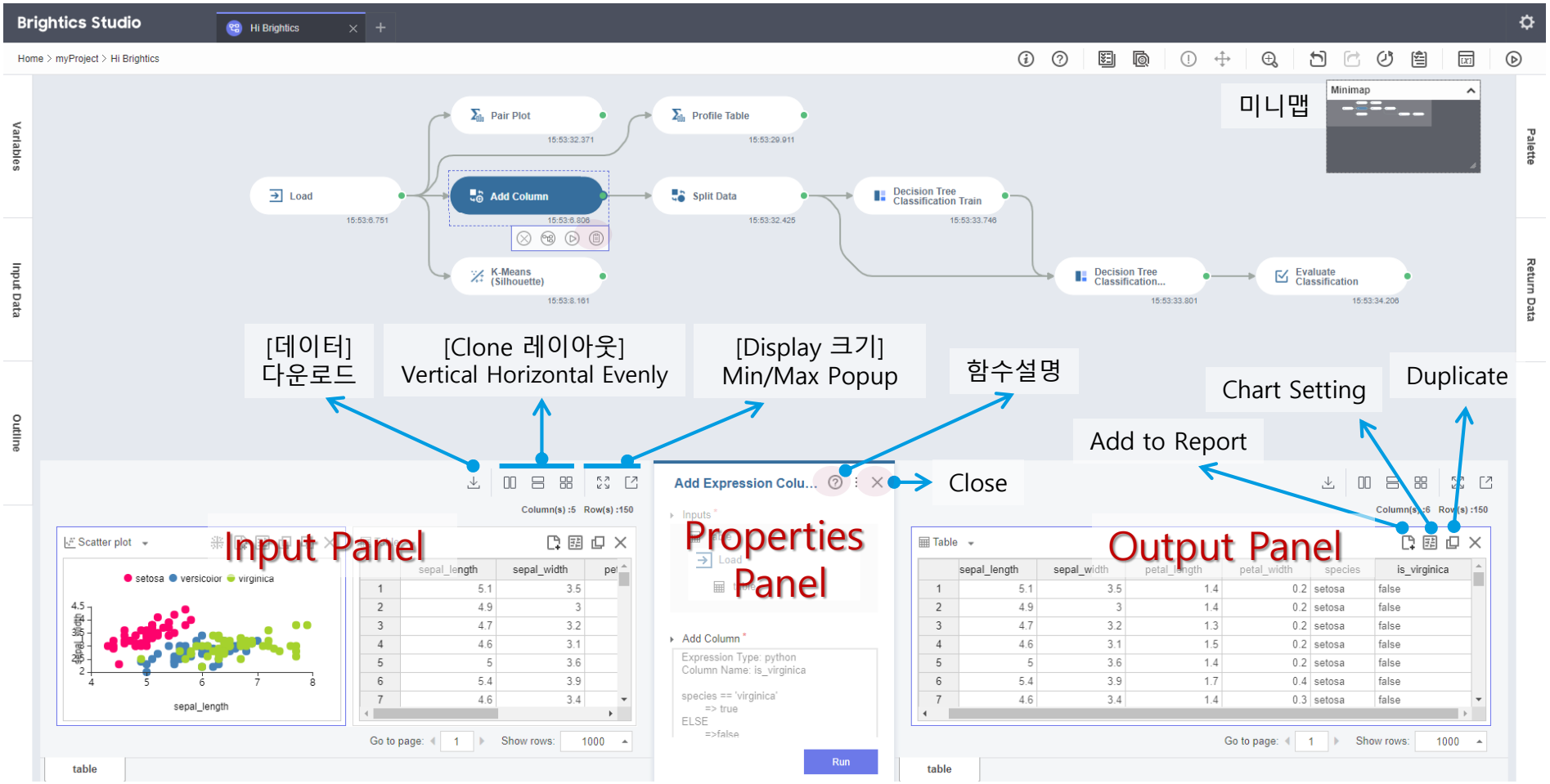


(용) 우측 날개창을 클릭하면 창이 열린다. 세 개의 탭이 있는데,
1. 함수를 보여주는 Funtion 탭,
2. 모델의 일부를 저장하는 Template,
3. 데이터를 외부로 보내고 받을 수 있는 Data 탭이다.

화면 구성 (3) - Model View(3), Sheet editor

Function Properties Panel - Parameter 입력 및 함수 실행

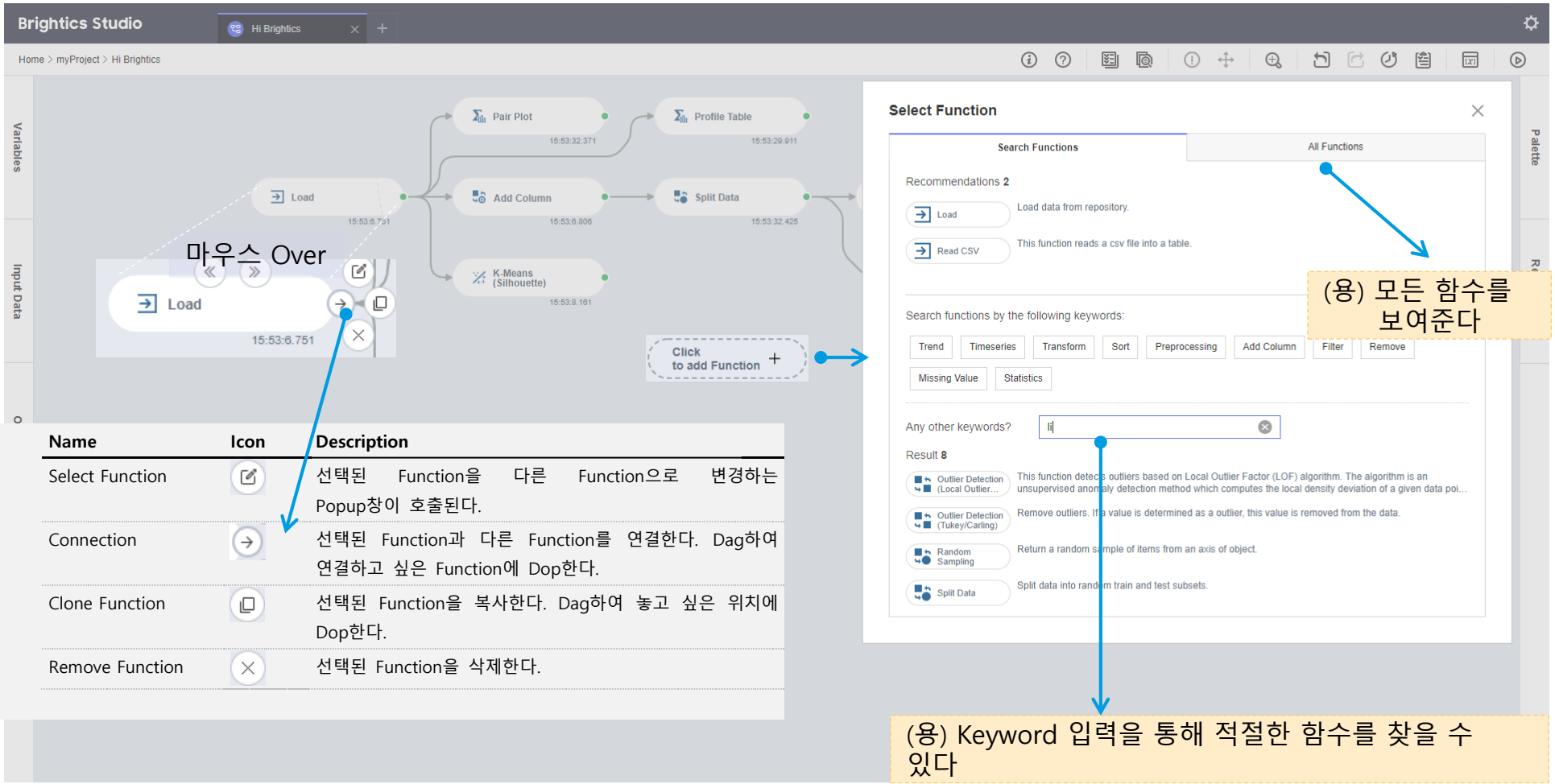
Data Panel - Function의 In/Out Data를 확인



화면 구성 (3) - Model View(4), Sheet editor

캔버스 : Data flow를 그리는 화면

Select Function : 캔버스에 함수를 입력

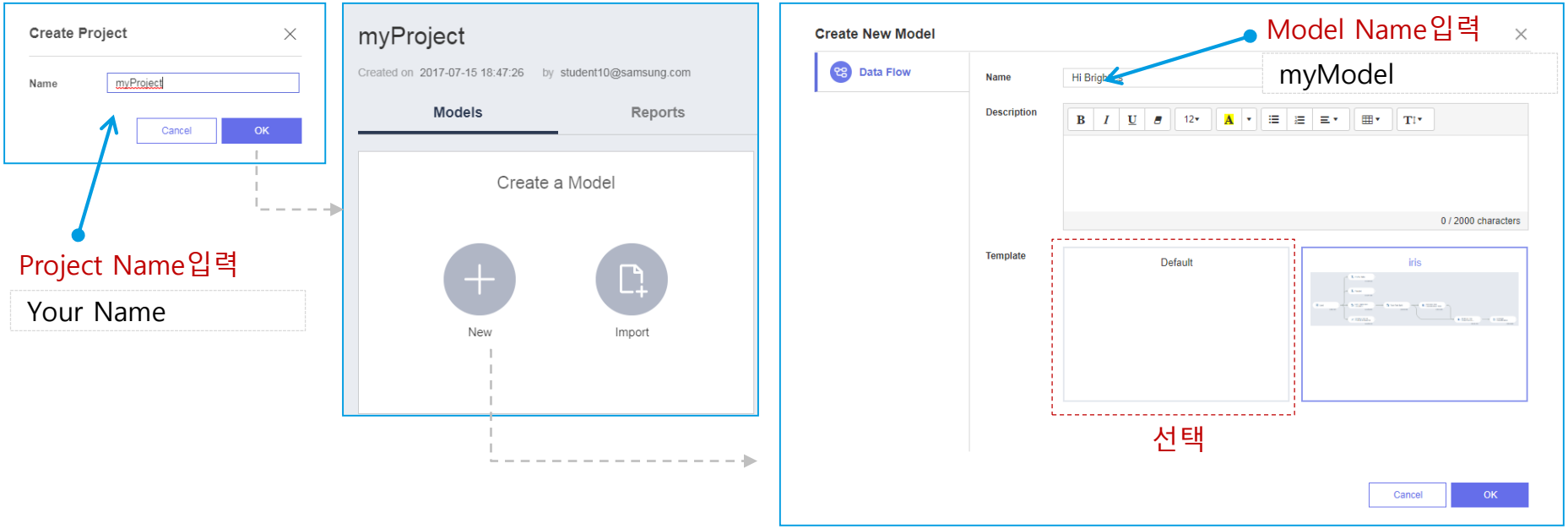


Name	Icon	Description
Select Function		선택된 Function을 다른 Function으로 변경하는 Popup창이 호출된다.
Connection		선택된 Function과 다른 Function를 연결한다. Dag하여 연결하고 싶은 Function에 Dop한다.
Clone Function		선택된 Function을 복사한다. Dag하여 놓고 싶은 위치에 Dop한다.
Remove Function		선택된 Function을 삭제한다.

Follow the Instruction

강사를 따라 하세요

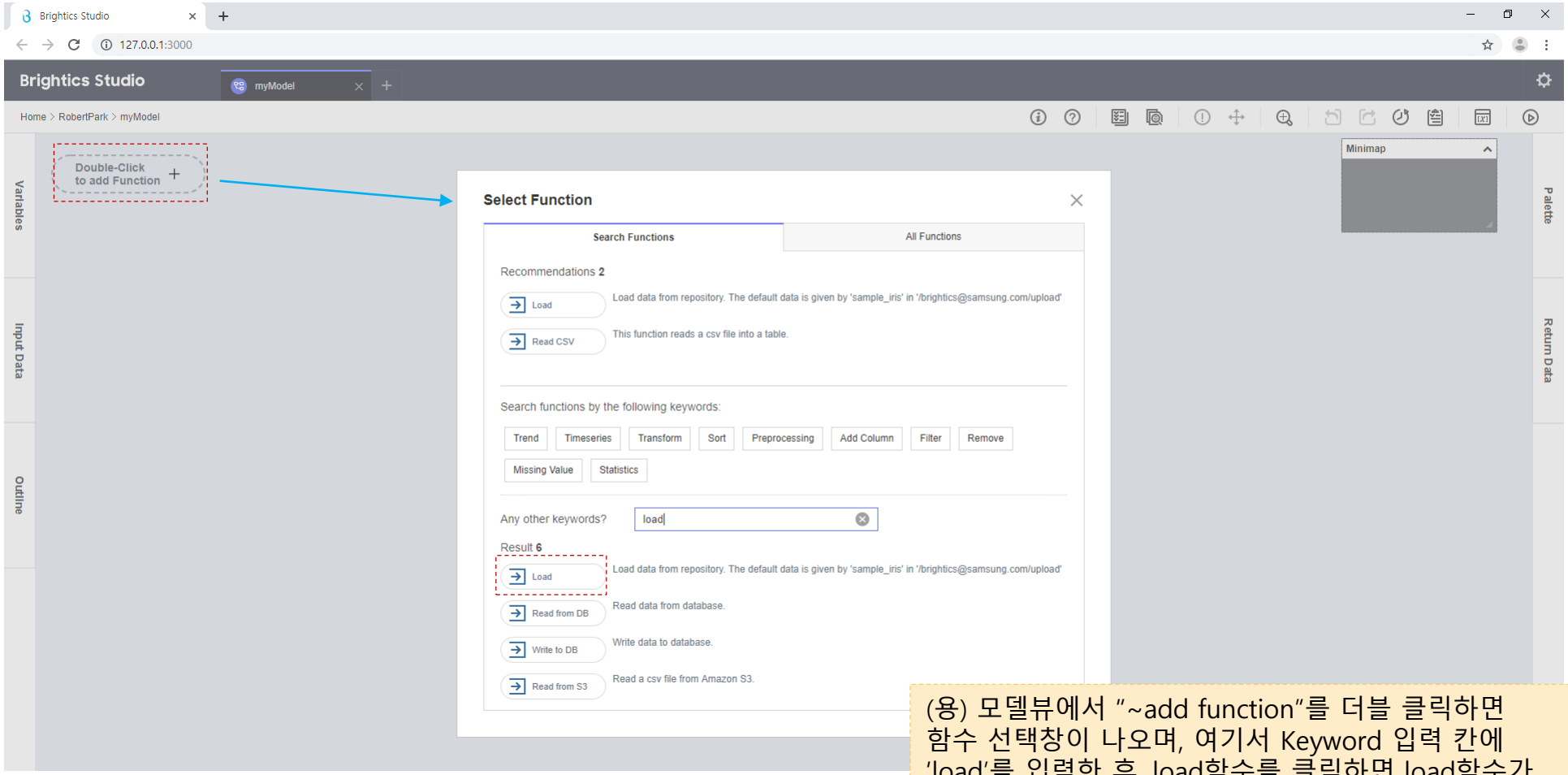
- 먼저 ownership을 가진 프로젝트와 모델을 생성합니다.



"CHROME 화면을 보세요"

Load data

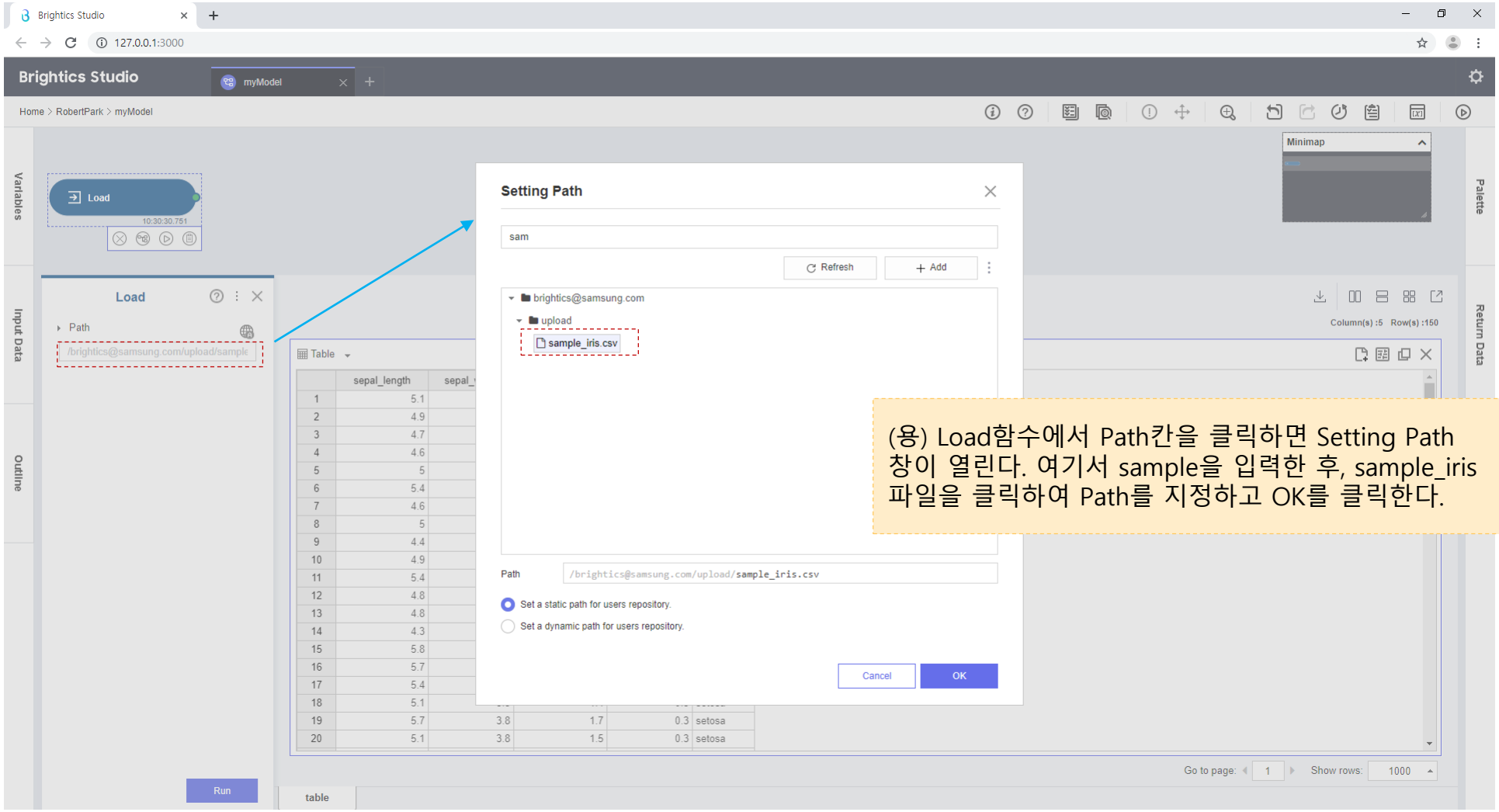
Load 함수를 배치한다.



(용) 모델뷰에서 "~add function"를 더블 클릭하면 함수 선택창이 나오며, 여기서 Keyword 입력 칸에 'load'를 입력한 후, load함수를 클릭하면 load함수가 캔버스에 배치된다.

Load data (sample_iris)

Setting Path 에서 sample_iris.csv 데이터 Load



(용) Load함수에서 Path칸을 클릭하면 Setting Path 창이 열린다. 여기서 sample을 입력한 후, sample_iris 파일을 클릭하여 Path를 지정하고 OK를 클릭한다.

2. Brightics Visual Analytics 기초 실습

- 시나리오#1 - 그래프
- 시나리오#2 - 데이터 입/출력
- 시나리오#3 - Load & Stat. Summary
- 시나리오#4 - 필터 (Filter)
- 시나리오#5 - 데이터 저장
- 시나리오#6 - 모델 복사/저장

실습 데이터 : iris.csv

자, 이제부터

강사를 따라 하면서 Brightics VA 실습을 하겠습니다.

집중해 주시기 바랍니다! 아니면, 길을 잃을 수 있습니다.

<Ronald Fisher>
1890~1962

Data set [edit]

Fisher's *Iris* Data

Sepal length	Sepal width	Petal length	Petal width	Species
5.1	3.5	1.4	0.2	*I. setosa*
4.9	3.0	1.4	0.2	*I. setosa*
4.7	3.2	1.3	0.2	*I. setosa*
4.6	3.1	1.5	0.2	*I. setosa*

Setosa

virginica

versicolor

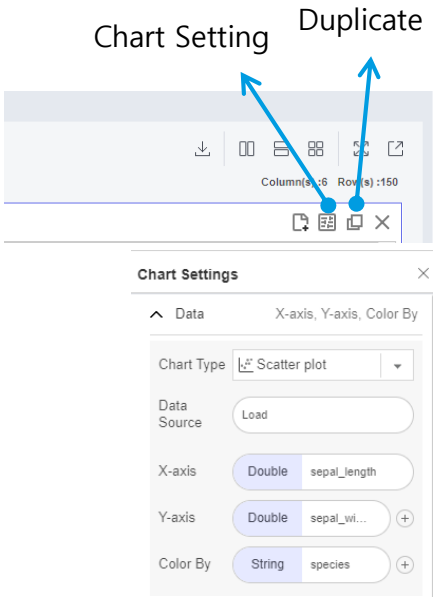
sepal — length, width

petal — length, width

시나리오#1 – 그래프

iris 데이터를 Repository에서 Load 하여, 그래프 작성

- **Model Editor에서 Repository데이터를 로딩한 후, 그래프를 그려본다.**
  1. Load함수 / Input path 클릭
  2. Repository에서 myiris.txt 파일 선택
  3. 메타데이터를 읽은 후, <RUN> 클릭
  4. 150개 data Row 확인



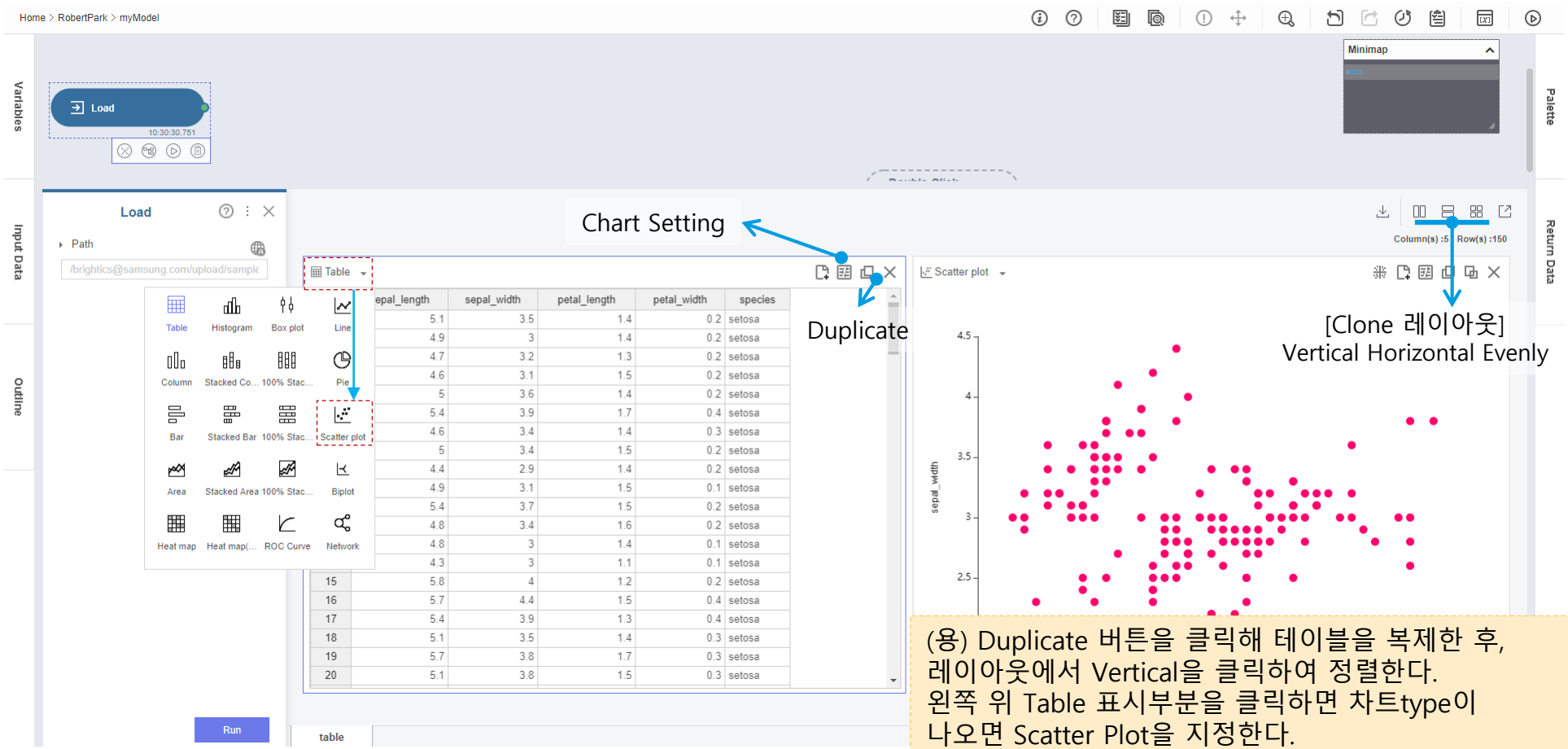
- **그래프 그리기**
  1. [Data panel] Chart type에서 Scatter Plot 선택
  2. Chart Setting icon 클릭
     1. Data, X=sepal_length, Y=sepal_width 선택
     2. Chart Setting 창 제거 (X 클릭)
  3. Duplicate icon 클릭
     1. Chart Setting icon 클릭
     2. Color By option에서 Species 선택

SepalLength와 SepalWidth간
가장 직선관계인 Species는?

그래프 분석

테이블을 차트로 변환한다. (scatter plot)

- duplicate를 클릭하여 차트창 복제
- chart setting으로 그래프 조정 (http://www.brightics.ai/docs/ai/s1.0/chart_user_guide/scatter)



차트세팅..데이터

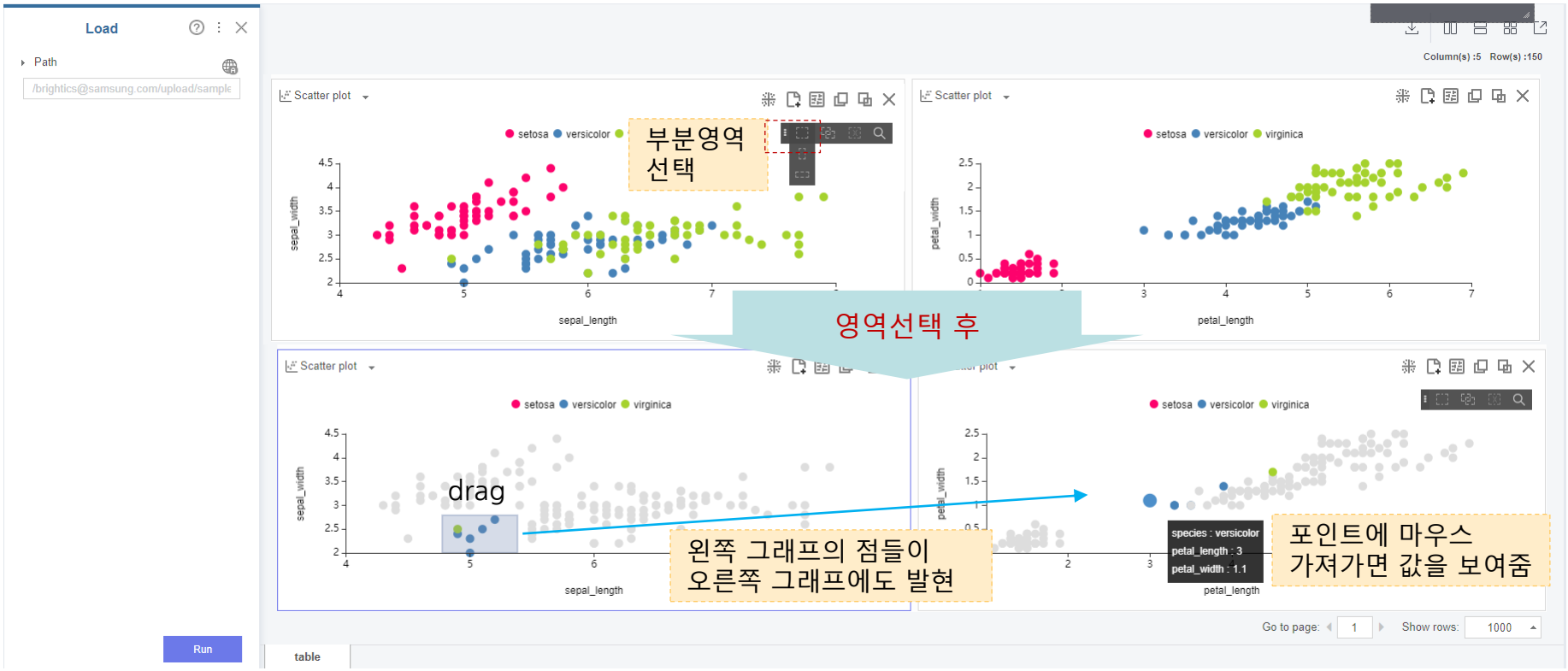
차트 조정

- 차트세팅/데이터 탭에서 color by 옵션을 species로 선택

그래프분석

두 개의 그래프로 연계 분석

- 오른쪽 그래프에서 x, y축을 petal_length, petal_width 로 변환 (Chart setting/ Data)
- 왼쪽그래프에서 부분영역 선택 (+자 마우스 드래그)
- 선택된 영역이 오른쪽 그래프에서 발현

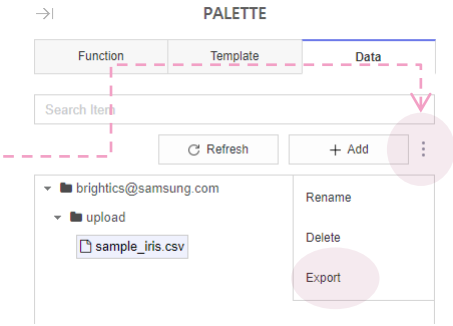


시나리오#2 - 데이터 입/출력

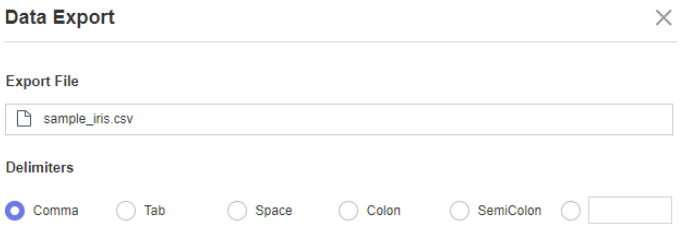
BR Repository의 데이터를 PC로 다운로드 하고, PC의 데이터를 BR로 업로드

*** Owner Proj/Model에서 작업할 것**

- **BR의 데이터를 PC로 다운로드**
  1. [BR] Palette / Data Tab에서 Upload Folder 열기
  2. [BR] sample_iris.csv 파일을 로컬PC로 Export (검색창에서 찾기)
  3. [PC] Iris.csv 파일을 폴더 열기하여 탐색기로 확인
  4. [PC] 파일 이름을 "myiris.csv"로 변경



- **PC의 데이터를 BR로 업로드**
  1. [BR] Pallette / Data Tab에서 <+Add> 클릭
     1. Local
     2. 'my_iris' 선택 <열기> 클릭
     3. Delimiter 선택 (Comma)
     4. Column name and type 선택 (숫자 double, 문자 string)
     5. <Finish> 클릭

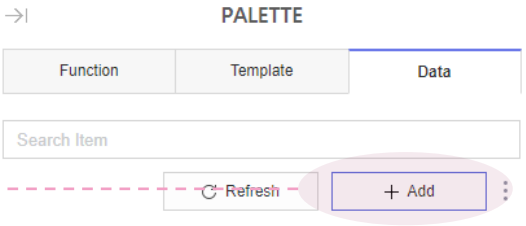


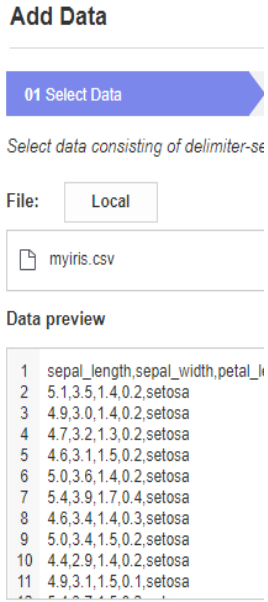
시나리오#2 - 데이터 입/출력

BR Repository의 데이터를 PC로 다운로드 하고, PC의 데이터를 BR로 업로드

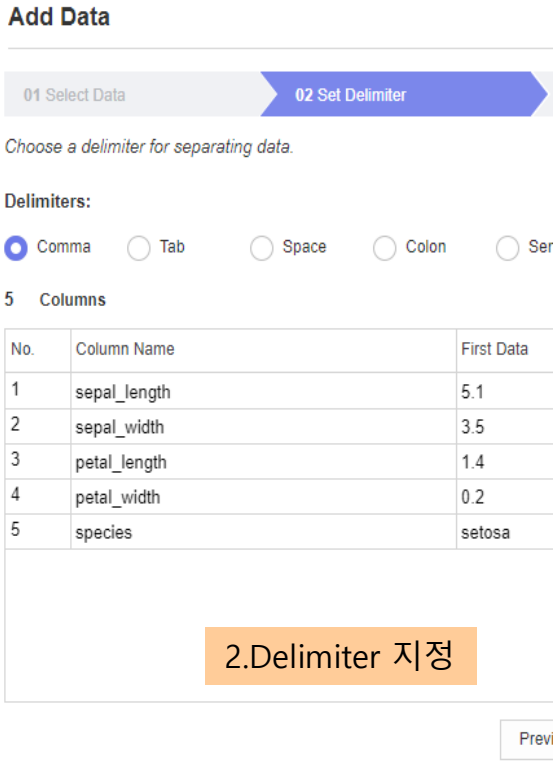
* Owner Proj/Model에서 작업할 것

- PC의 데이터를 BR로 업로드

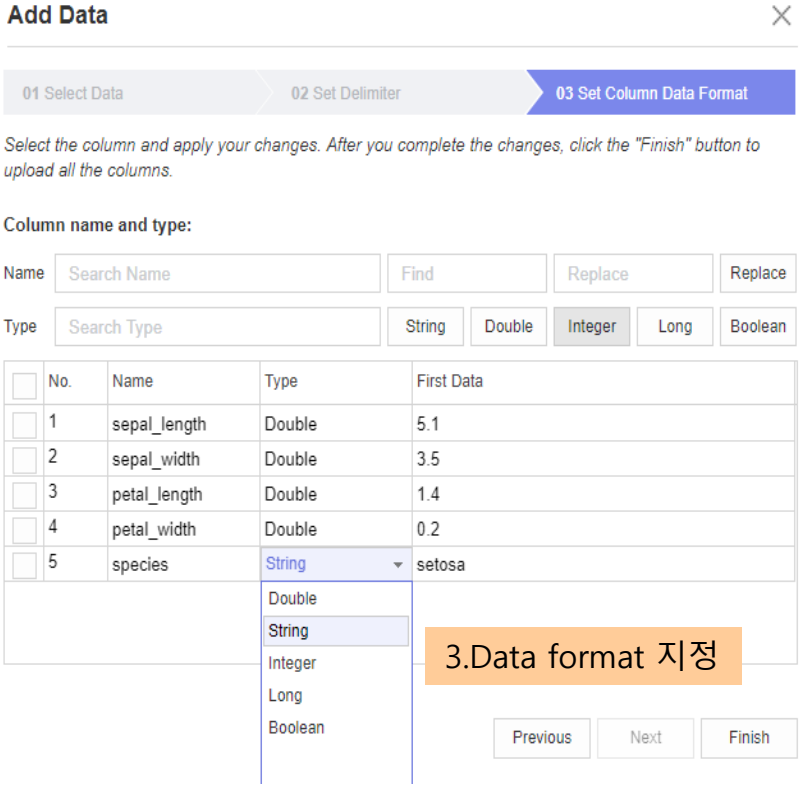




1.파일선택



2.Delimiter 지정

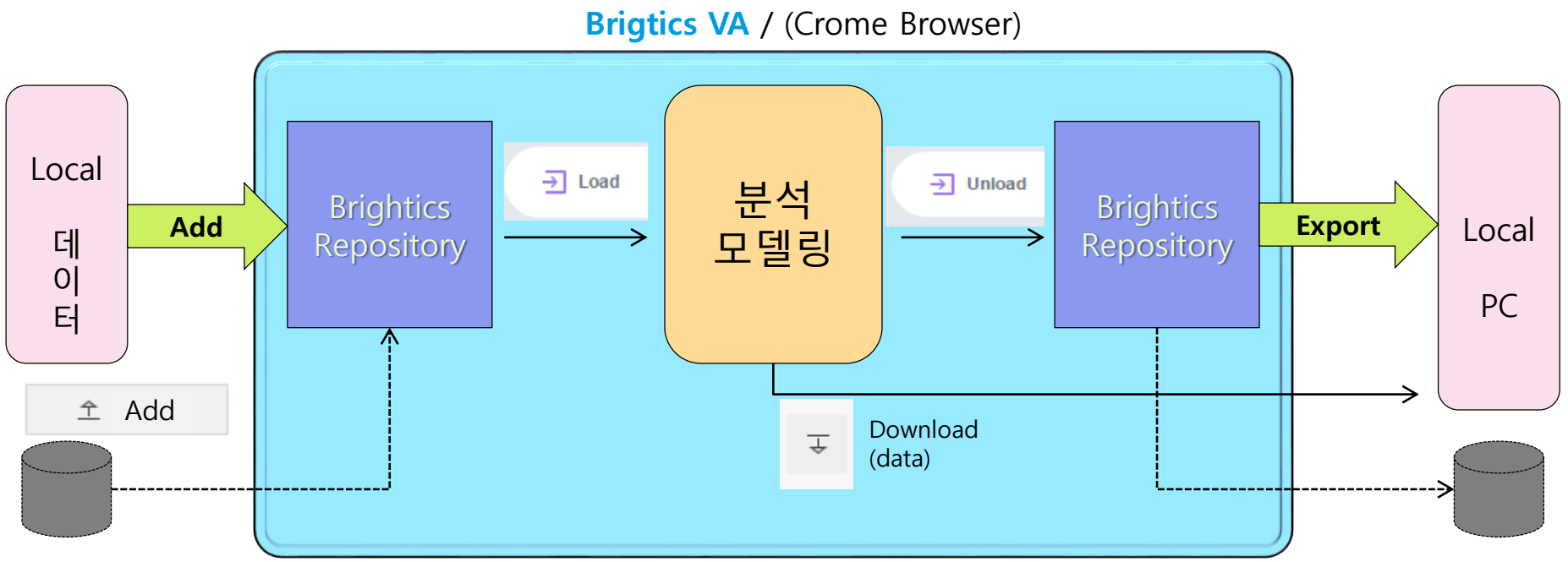


3.Data format 지정

참고. 데이터 Load/Unload/Add/Export

데이터 분석은 레파지토리에서 모델로 데이터를 로드(Load)하는 것으로 시작된다.

- 데이터를 Add할 때, Delimiter, Column type을 맞추는 작업 필요



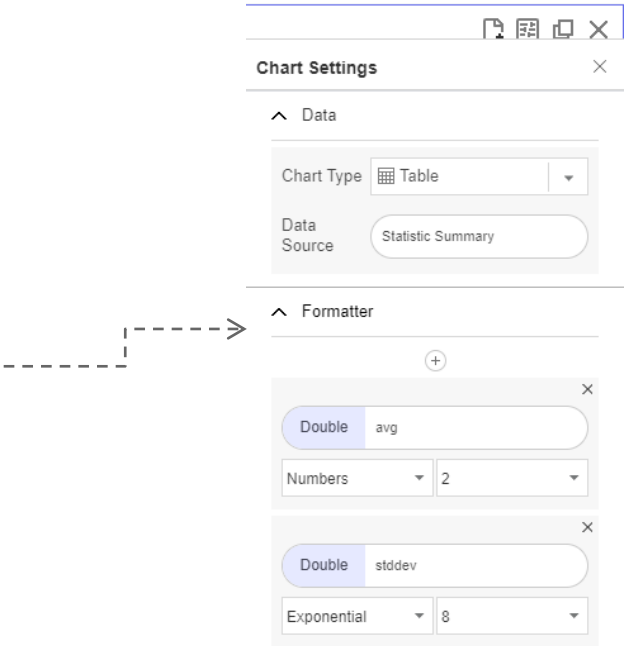
* Add, Export : 모델링 화면의 우측 Palette 날개창 / data 탭에서 선택

* Load, Unload: 모델링 함수에서 선택,

시나리오#3 – Load & Stat. Summary

Myiris 데이터를 Stat. Summary 함수로 평균, 표준편차를 계산

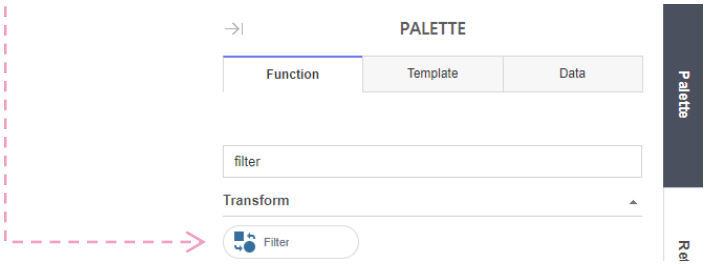
- **Model Editor에서 Stat. Summary 함수를 작성한 후, 통계값 계산**
  1. Load함수 옆에 (+) 마크가 나오게 한 후, 클릭
  2. Search Item 창에서 keyword 'summary' 입력
  3. Statistic Summary 아이콘 클릭
     1. Input Column에서 통계값을 계산할 대상을 선택
        - Select 클릭하여, SepalLength, Sepal Width 선택
     2. Target Statistic 에서 계산할 통계값 지정
        - Average, Standard Deviation 선택 <Run>
     3. 데이터칼럼 자릿수 Formatter 조정
        - Chart Setting // Formatter 클릭
        - Select Column에서 avg 선택 // Numbers 2 선택
        - (+) 클릭, Select Column에서 stddev 선택
        - Numbers 4 선택

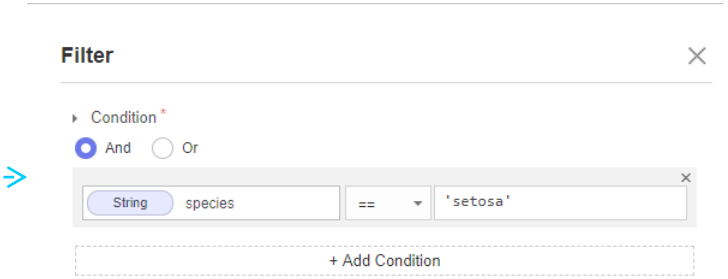


시나리오#4 - 필터 (Filter)

필터링 (Filter): Species가 'setosa'인 것만을 추출해 냄

- Model Editor에서 Filter 함수를 배치
  1. 오른쪽 날개 Palette창 열고, Function Tab 선택
  2. Search Item 창에서 'filter' 입력 후, 캔버스로 Filter 함수 Click & Drag_Down
     1. In Table: Load 함수에 connect
     2. Filter함수 Condition 지정
        - Condition창 클릭
        - Select Column 클릭 Species 선택
        - == 선택 (조건 지정)
        - 'setosa' 지정 (Value 지정) // OK
     3. <Run> 클릭
  3. Out Data Panel 확인
     1. 50개 data Row 확인

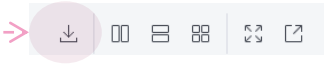


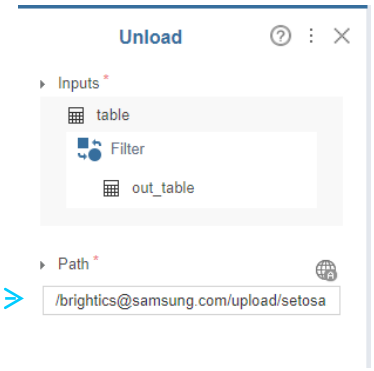


시나리오#5 – 데이터 저장

함수 결과(Setosa 데이터)를 Local PC/사용자 Repository에 저장

- Model Editor에 배치된 Filter 함수에서 데이터 다운로드
  - [to PC] Download Icon 클릭
    - 로컬 PC에 다운로드 (*함수이름*_out.csv) *#File Name 지정 가능*
  - [BR Repository]에 저장
    1. Unload 함수 배치
       1. Path 에 파일이름 입력
          - /brightics@samsung.com/upload/setosa.csv
          - Ok 클릭
       2. <Run> 클릭
  - 확인
    1. PALETTE // Data Tab
    2. 사용자 폴더 더블클릭





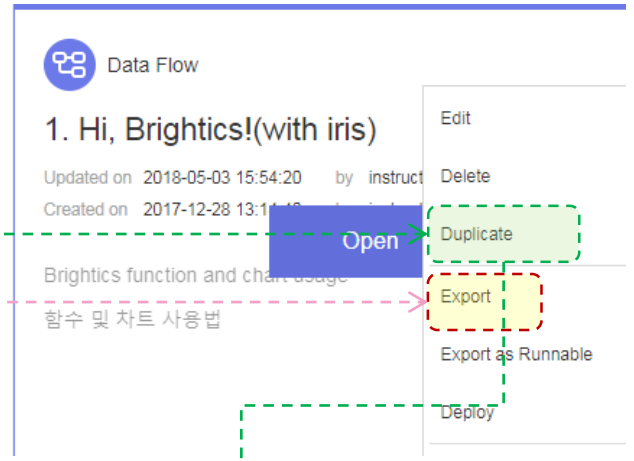
시나리오#6 – 모델 복사/저장

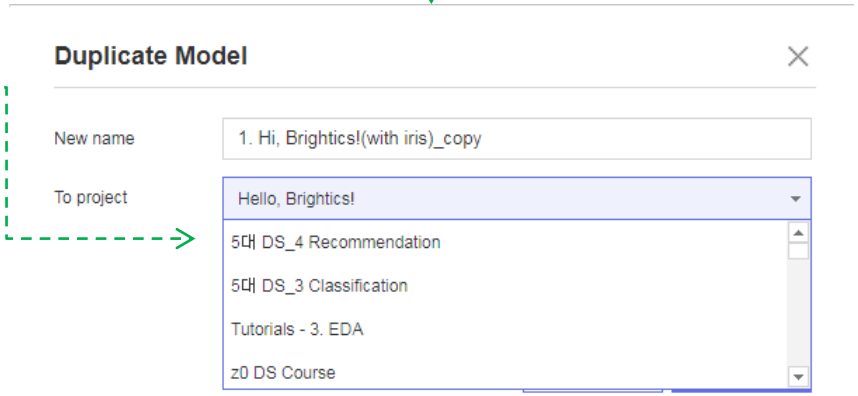
모델을 다른 프로젝트로 가져오기/ 모델을 내 PC에 저장하기

- **모델 → PC로 저장**
  1. Project View - Model List 에서 삼점(⋮)icon 선택
  2. Export 클릭 → <ok> 클릭
  3. 로컬 PC에 다운로드 (모델이름.json)

- **모델 → Project에 복사**
  1. Project View - Model List 에서 삼점(⋮) icon 선택
  2. Duplicate 클릭
     1. New Name 지정
     2. To Project 선택 → <ok> 클릭
     3. 선택된 Project로 복사

- **모델 import**





3. 시나리오 예제

- 종합시나리오#1 Toll Traffic
  https://cafe.naver.com/brighticsstudio/388
- 시나리오#2 K-car
  https://cafe.naver.com/brighticsstudio/389

*Toll Traffic 문제

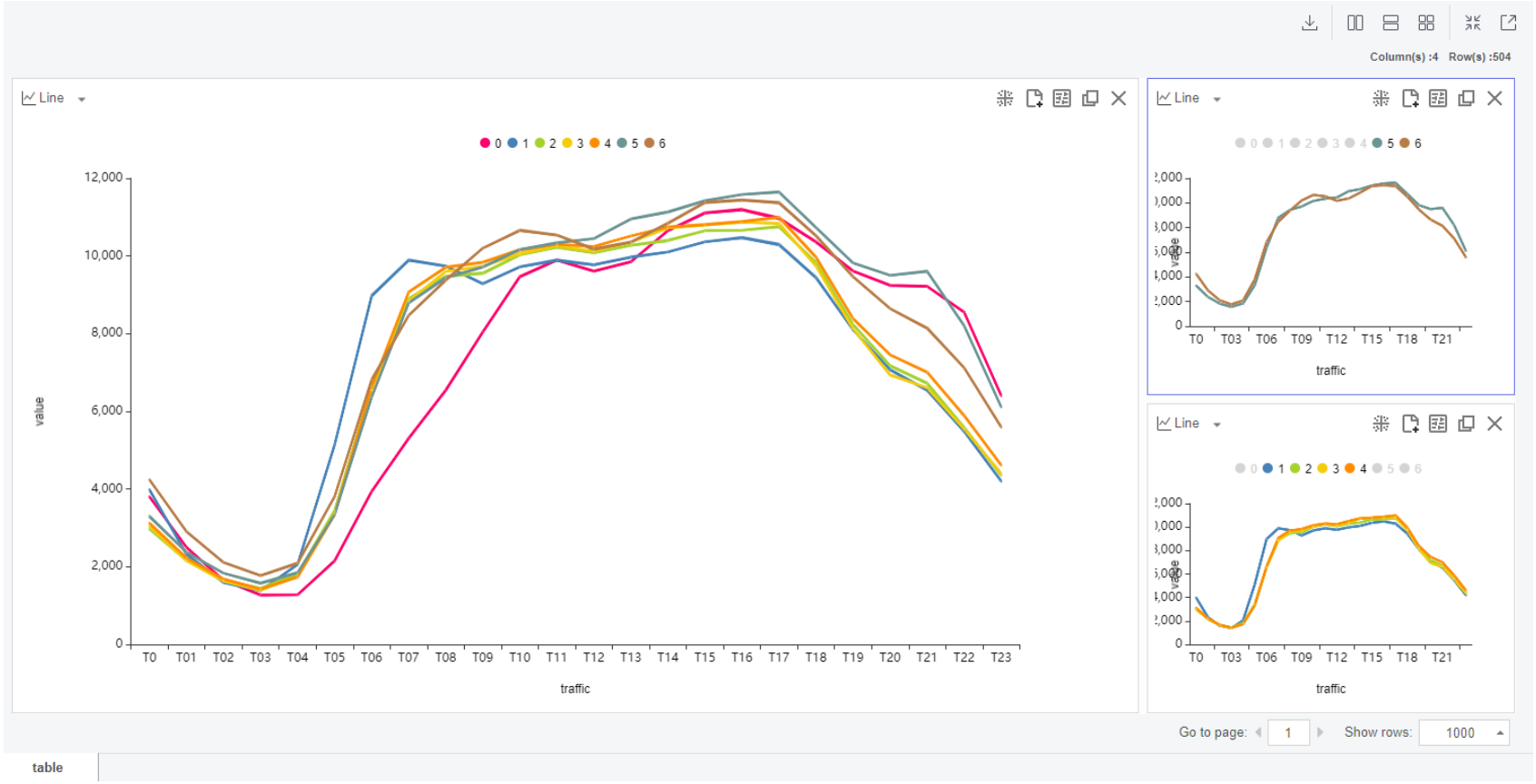
◉ 담당관의 고민: “징수원의 업무 강도를 효율적으로 관리하고 싶은 데, 요일 별 시간대별 특성은 어떠한가?” 몇 명이 적절할까?

- → 먼저 군집분석을 통해서 요일 별로 묶어 보자!! (when 2017.10.01)

분석 목표

◉ 시간대별 교통량을 기초로 요일 별 그룹화한다.

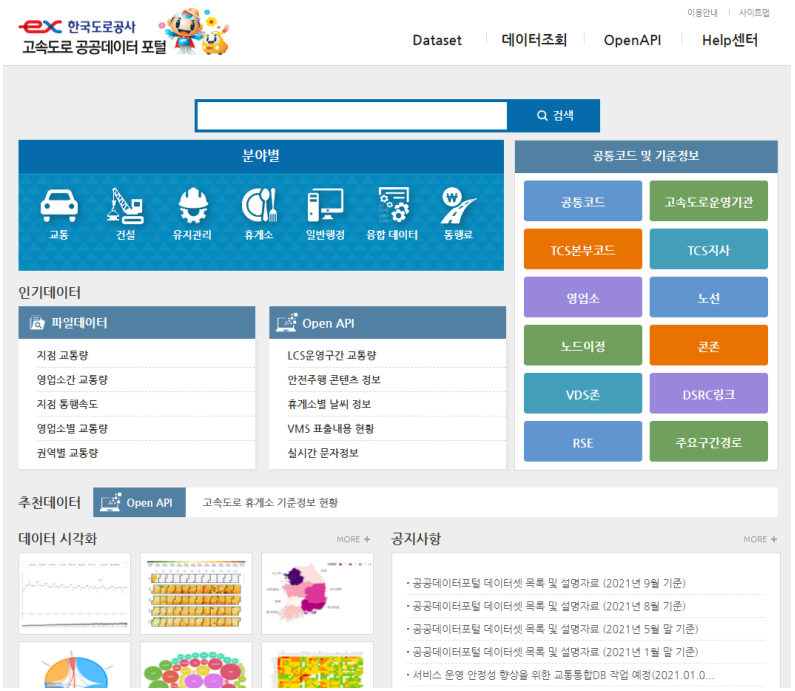
- 시각화 분석으로 요일 별 교통량 특성을 파악한다
- 군집분석을 통해 3개의 군집으로 그룹화한다.



데이터수집 및 준비

◉ 데이터 수집

- 도로공사의 데이터를 이용 데이터 정리 (.txt 파일)
- 사례를 위해 오산-동탄 구간의 데이터로 단순화
- Brightics로 데이터 +Add



Table

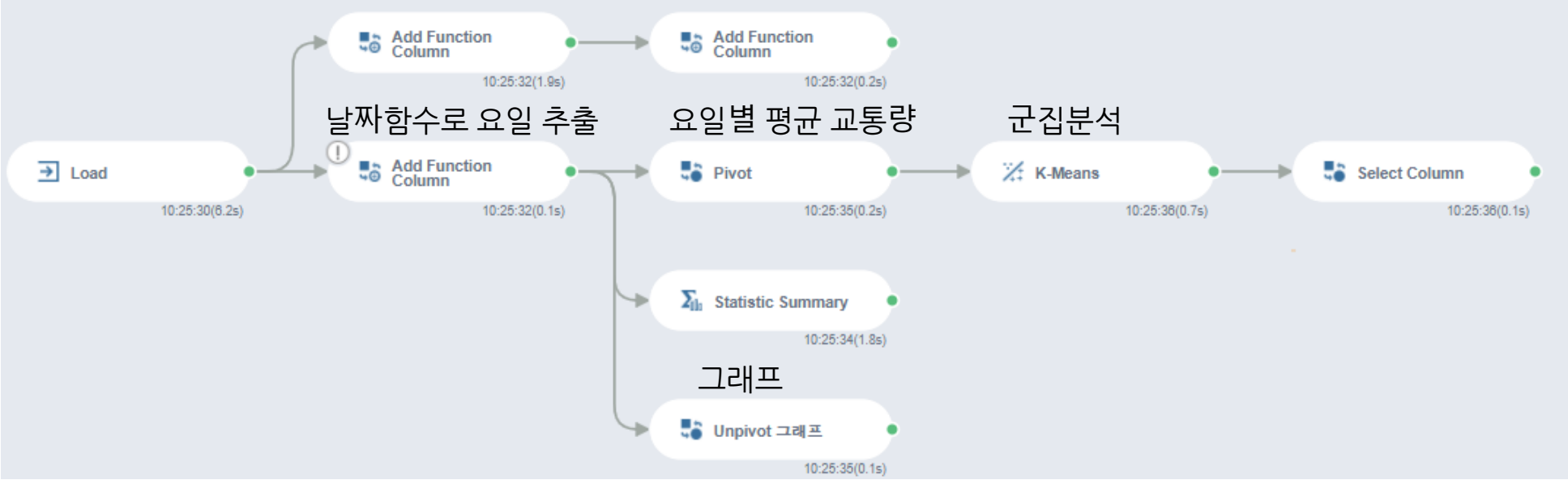
	date	T0	T01	T02	T03	T04
1	20170827	3795	2494	1649	1267	1278
2	20170828	3973	2321	1590	1378	2057
3	20170829	2966	2158	1653	1441	1812
4	20170830	3035	2150	1620	1385	1745
5	20170831	3118	2216	1674	1416	1724
6	20170901	3285	2372	1829	1568	1839
7	20170902	4234	2897	2102	1766	2082
8	20170903	3796	2499	1645	1263	1274
9	20170904	3980	2311	1585	1380	2052
10	20170905	2967	2155	1654	1446	1812
11	20170906	3029	2148	1626	1394	1740
12	20170907	3111	2219	1676	1425	1726
13	20170908	3279	2369	1832	1563	1846
14	20170909	4238	2902	2111	1766	2085
15	20170910	3812	2509	1664	1287	1288
16	20170911	3995	2335	1615	1396	2067
17	20170912	2984	2178	1669	1451	1833

traffic_Osan_Dongtan.txt

데이터 분석 및 모델링

◉ Brightics 를 통해 데이터 탐색 및 분석

- 전처리로 요일 추출
- 데이터 탐색을 통해 전반적인 데이터 파악
- 군집분석모델링으로 요일별 그룹화

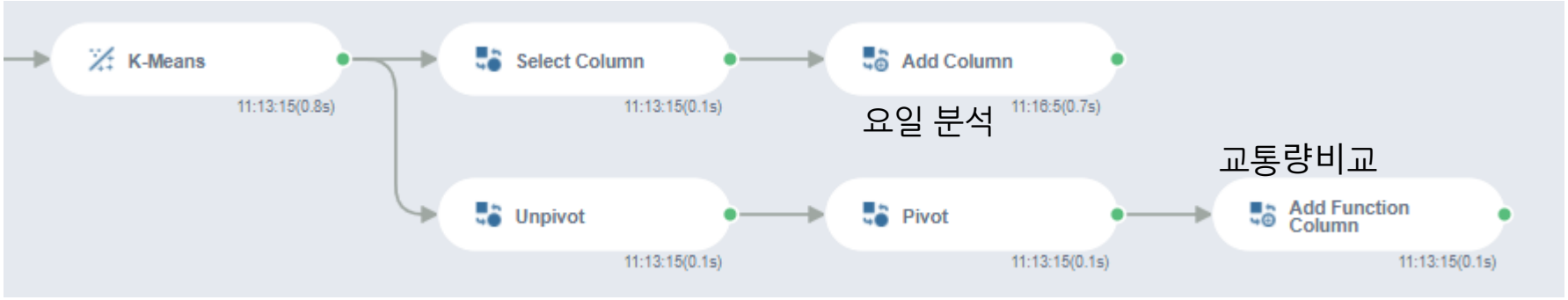


Toll Traffice.json

추가 분석 및 결론

◉ Brightics 를 통해 추가 분석

- 금요일과 교통량 패턴이 유사한 요일은 다음 중 무엇인가?
- 다른 요일과 교통량이 가장 독특한 요일은 다음 중 무엇인가?
- 금요일 그룹의 교통량은 월요일 그룹의 교통량의 몇%인가?



Toll Traffice.json

*Kcar-used car 중고차 선택 문제

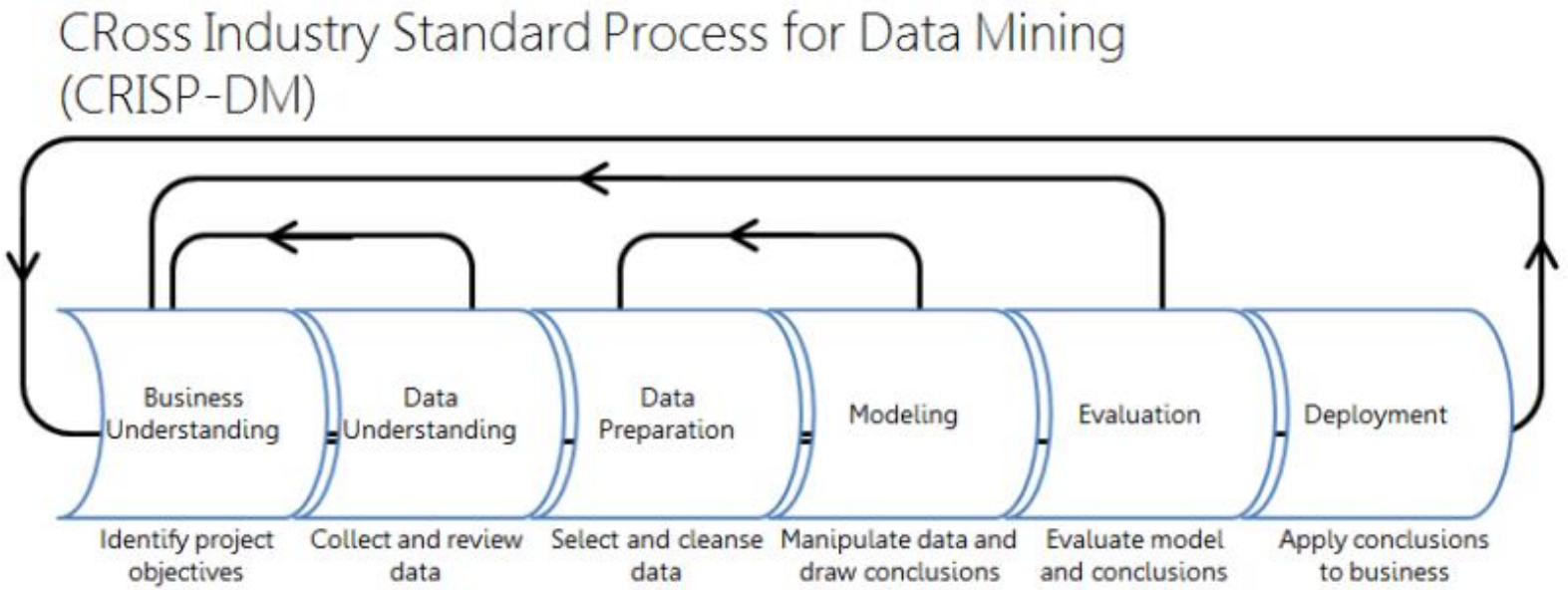
◉ 고프로의 고민: "차를 사야 하는데, 가장 비용효과적인 중고차는 무엇일까?"

→ 회귀분석을 통해서 가장 경제적인 자동차를 선택해 보자!! (when 2019.04.01)

*본 모델에서 특정회사의 홍보의도는 전혀 없습니다 ~!!~

분석 절차

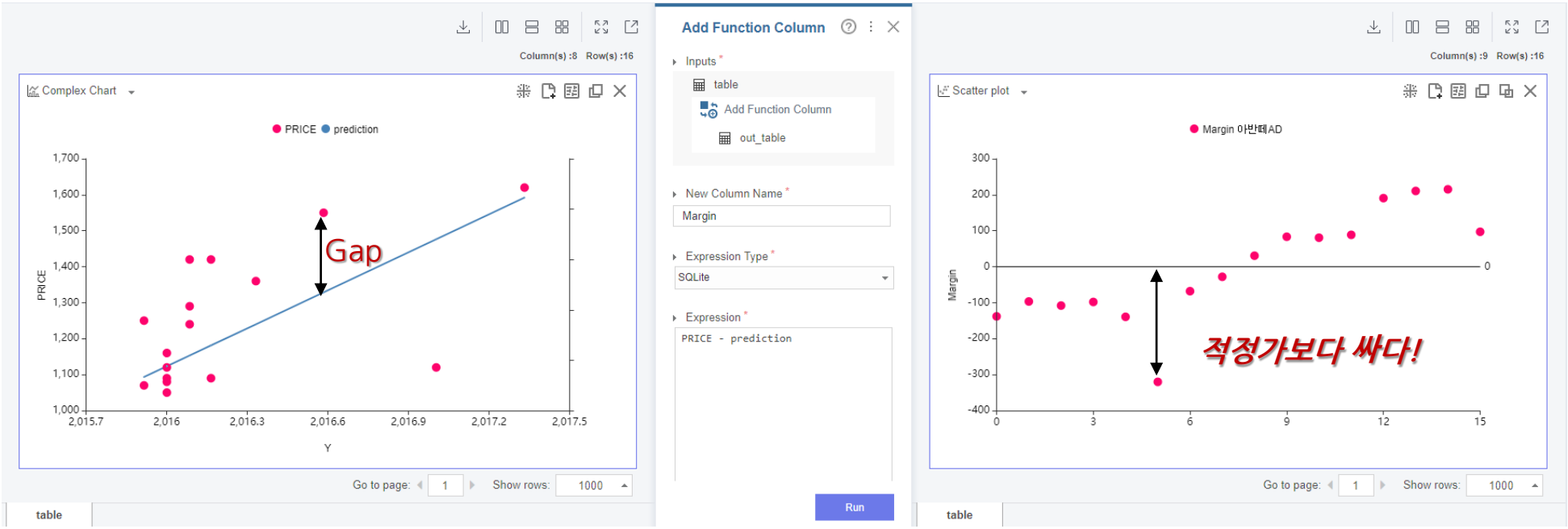
- 분석 절차:
- 분석목표 설정 → 데이터 수집 → 데이터 준비 → EDA(탐색) → 데이터 분석 … → 결론



분석 목표

◉ 가장 비용효과적인 중고차를 선택한다.

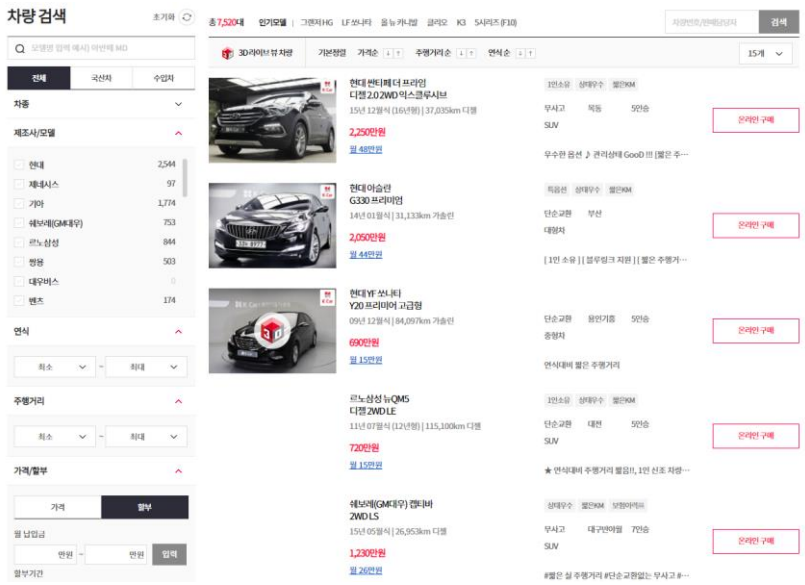
- 회귀분석을 통해, 중고차의 적정가격을 산출
- 판매가격과 적정가격과의 차이(Margin or Gap)를 계산
- Margin이 가장 작은 차량을 도출

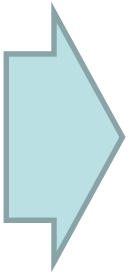


데이터수집 및 준비

◉ 데이터 수집

- 공신력 있는 온라인매장에서 데이터 수집 (python - github)
- 엑셀을 이용하여 데이터 정리 (.csv파일)
- Brightics로 데이터 +Add





Table

	NAME	YEAR	KM	PRICE	FUEL	AGE
1	아반떼AD	2016-01-01	99249	1050	디젤	1186
2	아반떼AD	2015-12-01	75806	1070	가솔린	1217
3	아반떼AD	2016-01-01	22335	1080	가솔린	1186
4	아반떼AD	2016-01-01	87339	1090	디젤	1186
5	아반떼AD	2016-03-01	77861	1090	디젤	1126
6	아반떼AD	2017-01-01	94000	1120	디젤	820
7	아반떼AD	2016-01-01	26247	1120	가솔린	1186
8	아반떼AD	2016-01-01	70538	1160	디젤	1186
9	아반떼AD	2016-02-01	44348	1240	가솔린	1155
10	아반떼AD	2015-12-01	39485	1250	가솔린	1217
11	아반떼AD	2016-02-01	25173	1290	가솔린	1155
12	아반떼AD	2016-05-01	40231	1360	가솔린	1065
13	아반떼AD	2016-03-01	48821	1420	가솔린	1126
14	아반떼AD	2016-02-01	31251	1420	가솔린	1155
15	아반떼AD	2016-08-01	11705	1550	가솔린	973

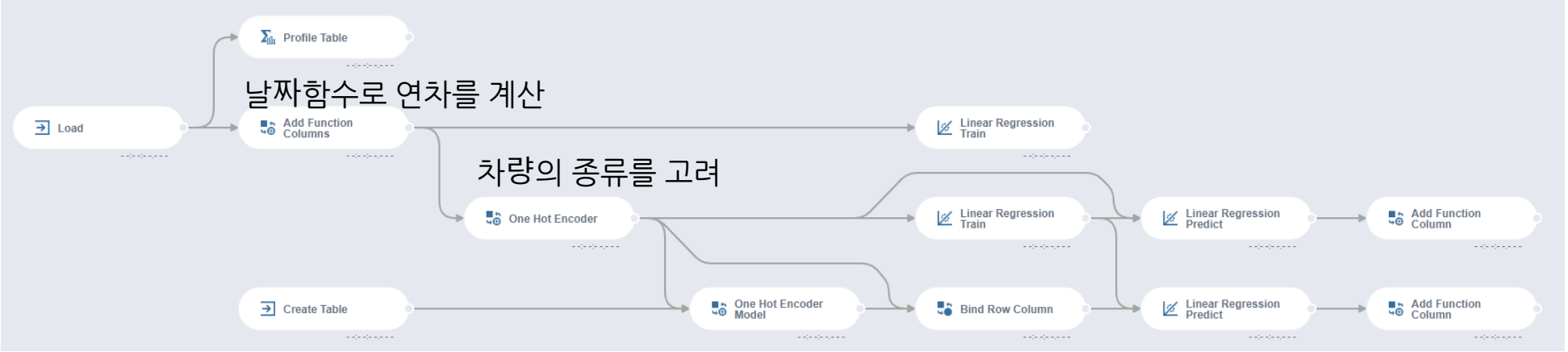
table

zoonggo.csv

데이터 분석 및 모델링

◉ Brightics 를 통해 데이터 분석 및 모델링

- 데이터 탐색을 통해 전반적인 추세 파악
- 회귀분석모델링으로 적정가격 구축
- 회귀분석 모델을 Refine

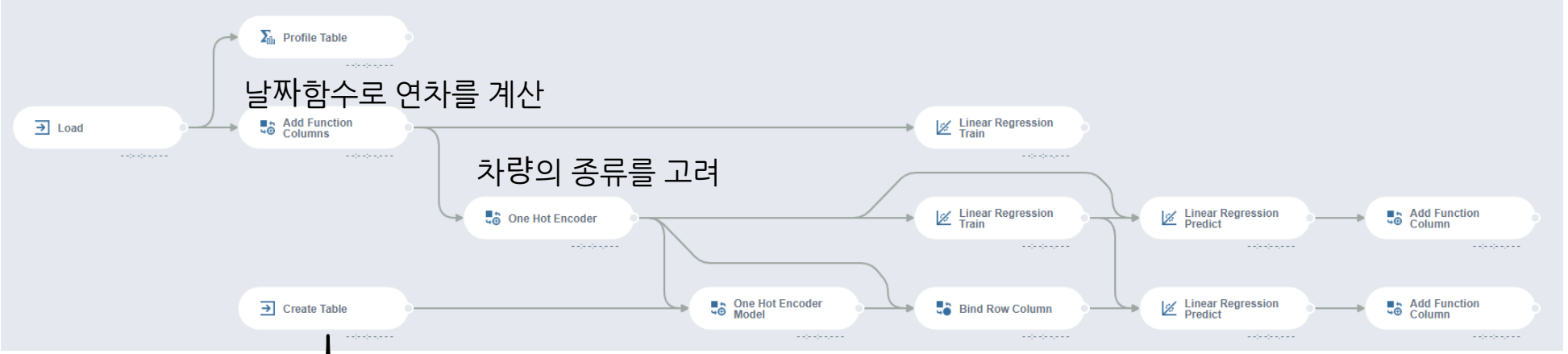


Kcar_Usedcar.json

추가 분석 및 결론

◉ Brightics 를 통해 데이터 분석 및 모델링

- What about Brand New Car?
- Overestimate or Underestimate?

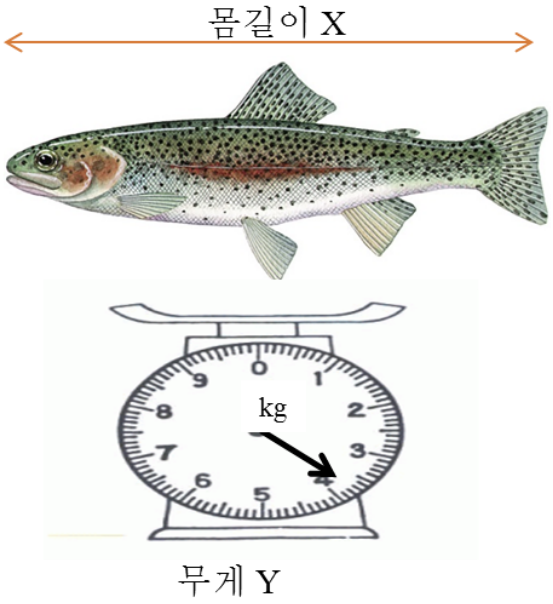


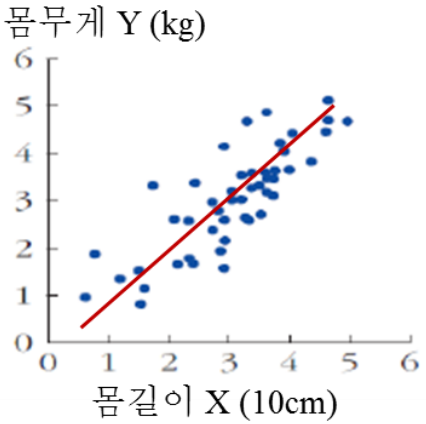
새 차인 경우는 어떠한가?

	NAME	YEAR	KM	PRICE	FUEL	AGE	MKM
1	아반떼AD	2019	4	1680	가솔린	0.1	0.1

참고. About Linear Regression

- 회귀 분석
  - 한 변수를 다른 변수(들)의 함수 관계로 표현하는 것





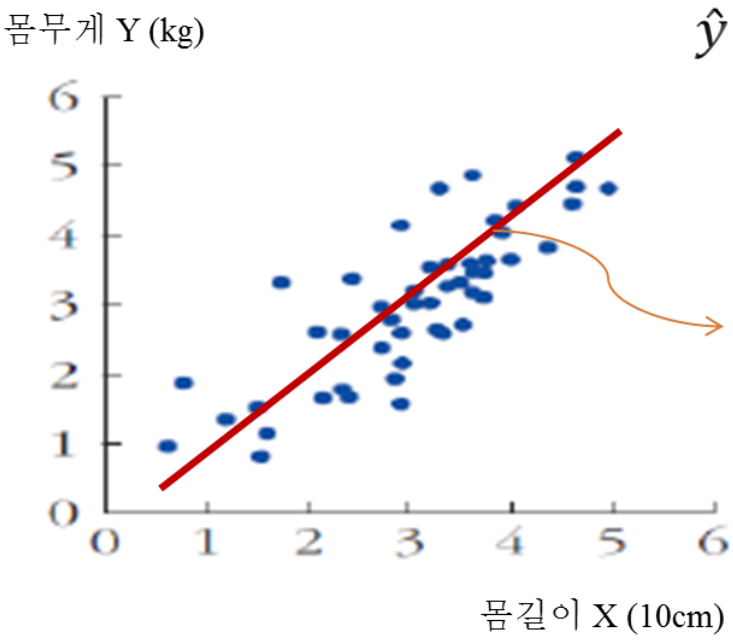
(용) 낚시를 좋아하는 홍만이가 어느 날 송어를 많이 잡았습니다. 다양한 크기의 송어를 잡았지요.

그런데, 송어의 몸길이와 무게간에 서로 어떤 관계가 있을까요? 물론, 살찐 송어도 있고 날씬한(?) 송어도 있겠지만 일반적으로 비례관계가 있을 겁니다.

이를 '상관관계가 있다'라고 하며, 이 관계를 함수로 표현한 것이 회귀분석입니다.

참고. About Linear Regression

- 종속변수 Y를 독립변수 X로 설명할 때, 이 관계가 선형(직선) 관계인 경우 선형회귀분석이라고 함.
- Y=f(X), Y=b0+ b1*X (y=a+b1*x1+b2*x2 처럼 여러 독립변수도 가능)



$$\hat{y} = b_0 + b_1 x$$

$$b_1 = \frac{\sum(x-\bar{x})(y-\bar{y})}{\sum(x-x)^2}$$

$$b_0 = \bar{y} - b_1\bar{x}$$

y (kg) = 1.12 x - 0.2

X: 송어의 몸길이 (10cm)
Y: 송어의 무게 (kg)

(용) 앞서, 송어의 길이와 몸무게에 상관관계가 있다고 살펴보았는데,

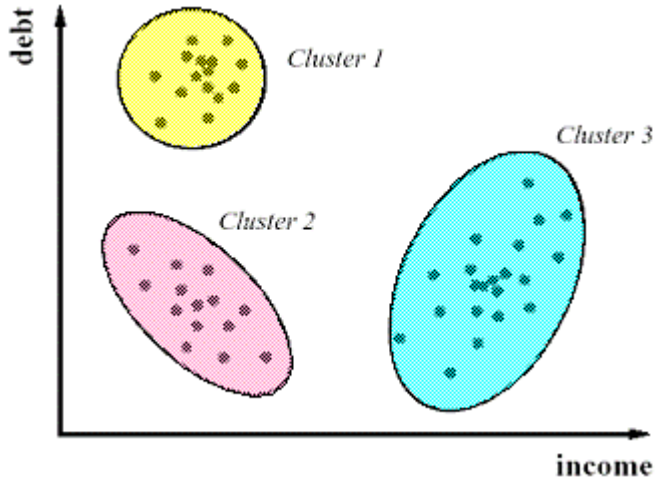
그렇다면, 송어의 몸무게와 몸길이 간에 어떤 함수 관계가 있는지를 파악하는 것이 회귀분석입니다.

일차함수 y = ax + b 의 형태에서 a와 b값을 추정하는 것입니다.

참고. K-means Clustering

전체 데이터를 K개의 군집으로 구별해 주는 것.

K means Clustering



공기놀이 하듯, 가까운 거리의 점들끼리 하나로 묶어주는 것. 그 묶음이 K개

n개의 d-차원 데이터 오브젝트 ($\mathbf{x}_1$, $\mathbf{x}_2$, …, $\mathbf{x}_n$) 집합이 주어졌을 때, K-평균 알고리즘은 n개의 데이터 오브젝트들을 각 집합 내 오브젝트 간 응집도를 최대로

하는 개의 집합 $\mathbf{S} = \{S_1, S_2, \ldots, S_k\}$ 으로 분할한다.
다시 말해, $\boldsymbol{\mu}_i$가 집합 S_i의 중심점이라 할 때

각 집합별 중심점~집합 내 오브젝트간 거리의 제곱합을 최소로 하는 집합 **S**를 찾는 것이 이 알고리즘의 목표다. 이 목적 함수의 전역 최솟값 (global minimum) 을 찾는 것은 NP-난해 문제이므로, 언덕 오르기 (hill climbing) 방식으로 목적 함수의 오차를 줄여나가며 지역 최솟값 (local minimum) 을 발견했을 때 알고리즘을 종료함으로써 근사 최적해를 구한다.