

## Chapter. 01

[ 비주얼코딩 ] 코딩없이 하는 데이터 분석 SAS

# 103. 데이터 탐색

FAST CAMPUS  
ONLINE

직장인을 위한 데이터 분석

강사. 최윤진

## Chapter. 06

# 코딩 없이 하는 데이터 분석 SAS

## 03. 데이터 탐색

# I 데이터

<http://bit.ly/vc-sas-a>  
d

변수	변수 설명
TV	TV 광고비
radio	라디오 광고비
newspaper	신문 광고비
sales	매출액

## I 데이터 탐색

## ▼ 데이터 탐색

```
[ ] # 라이브러리를 불러옵니다.
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns

# 데이터를 불러옵니다.
from google.colab import files
uploaded = files.upload()
for fn in uploaded.keys():
    print('파일을 불러왔습니다 "{name}" {length} bytes'.format(
        name=fn, length=len(uploaded[fn])))
```

파일 선택 선택된 파일 없음

Upload widget is only available when the cell has been executed in the current browser session. Please rerun this cell to enable.

Saving Advertising.csv to Advertising.csv  
파일을 불러왔습니다 "Advertising.csv" 4756 bytes

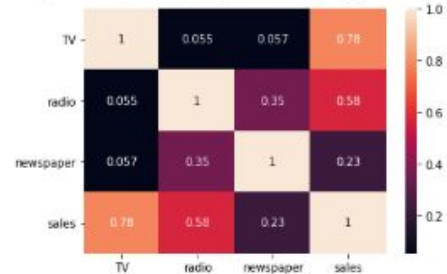
```
[ ] df = pd.read_csv("Advertising.csv")
print(df.shape)
df.head(10)
```

(200, 5)

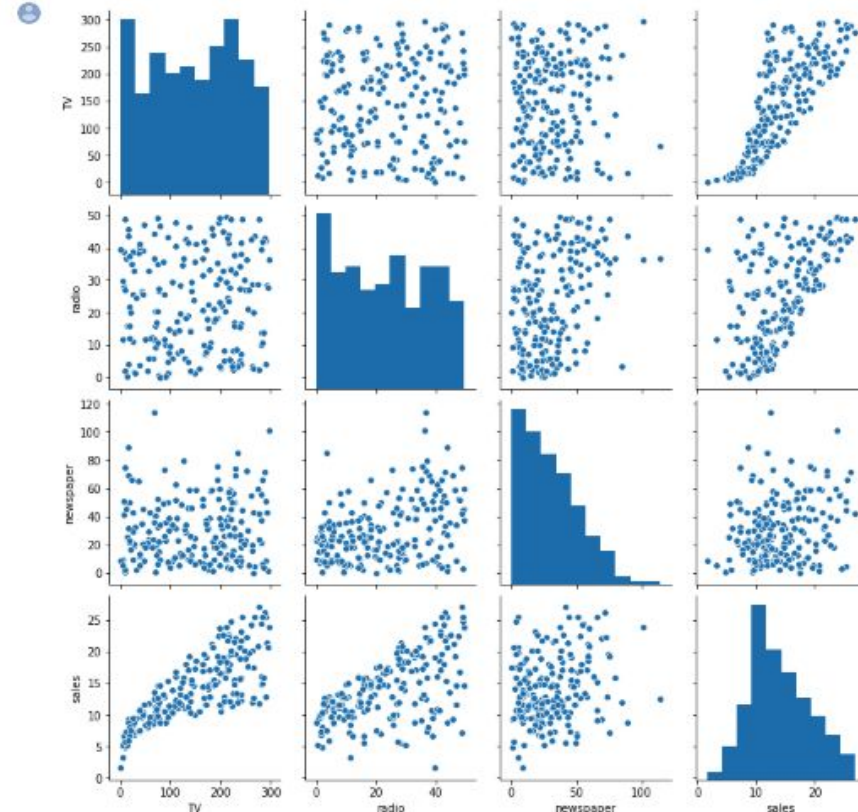
	Unnamed: 0	TV	radio	newspaper	sales
0	1	230.1	37.8	69.2	22.1
1	2	44.5	39.3	45.1	10.4
2	3	17.2	45.9	69.3	9.3
3	4	151.5	41.3	58.5	18.5
4	5	180.8	10.8	58.4	12.9
5	6	8.7	48.9	75.0	7.2
6	7	57.5	32.8	23.5	11.8
7	8	120.2	19.6	11.6	13.2
8	9	8.6	2.1	1.0	4.8
9	10	199.8	2.6	21.2	10.6

```
[ ] sns.heatmap(corr,annot=True)
```

<matplotlib.axes.\_subplots.AxesSubplot at 0x7fcaae31ef98>



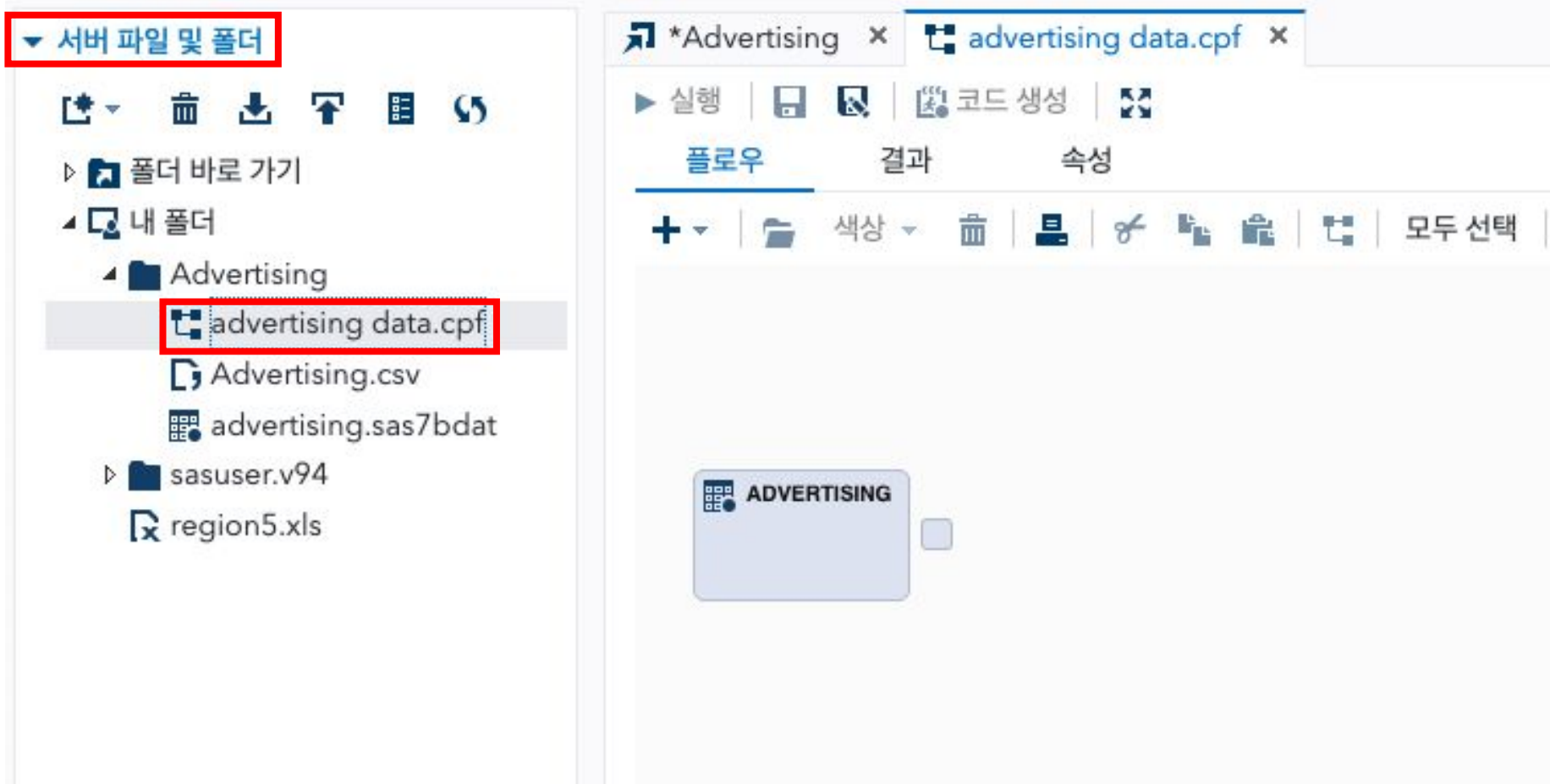
```
[ ] sns.pairplot(df[["TV", "radio", "newspaper", "sales"]])
plt.show()
```



# I 프로세스 플로우 불러오기

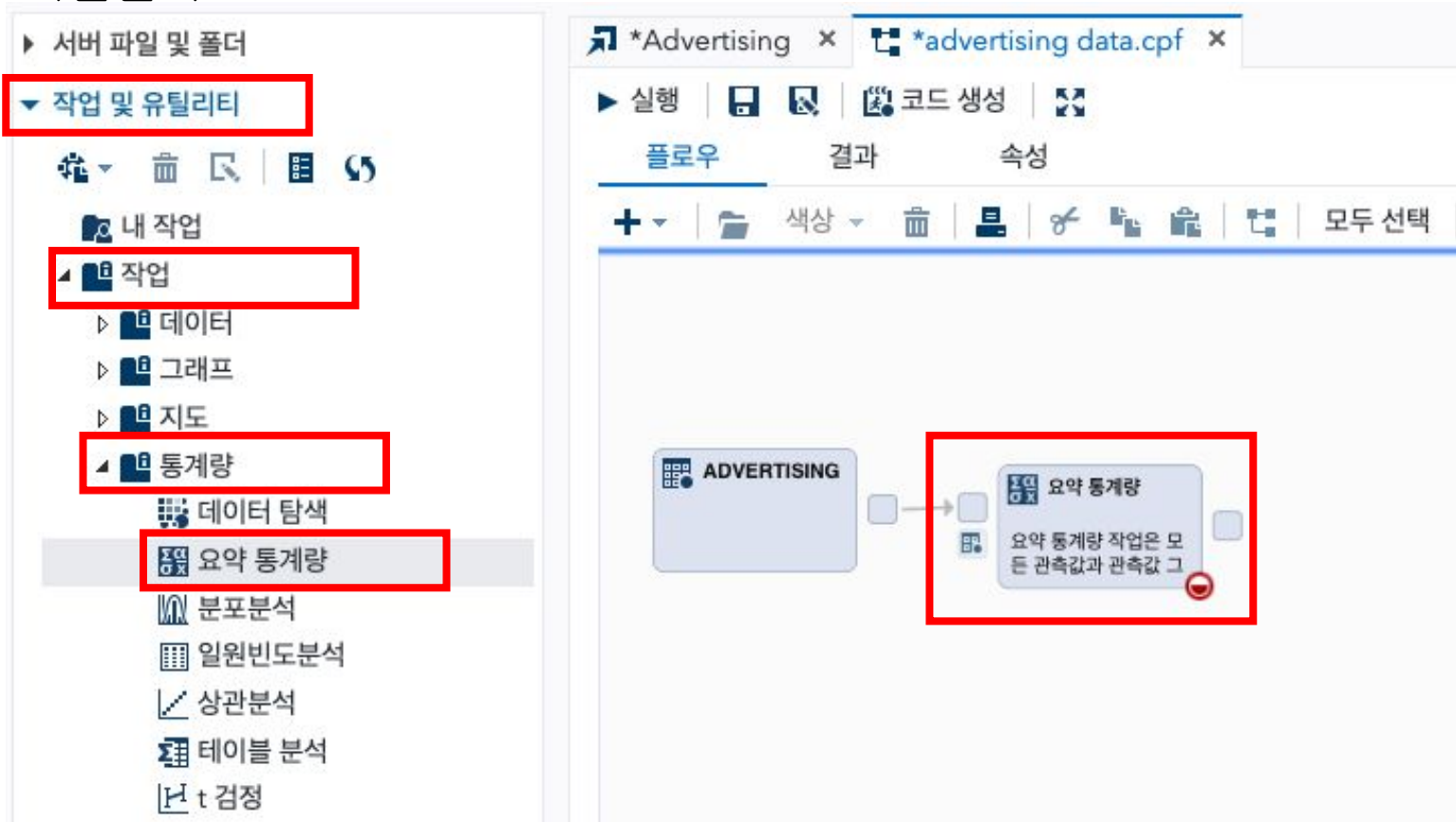
서버 파일 및 폴더 > 내 폴더 > Advertising > advertising data.cpf

실행



# I 요약 통계량

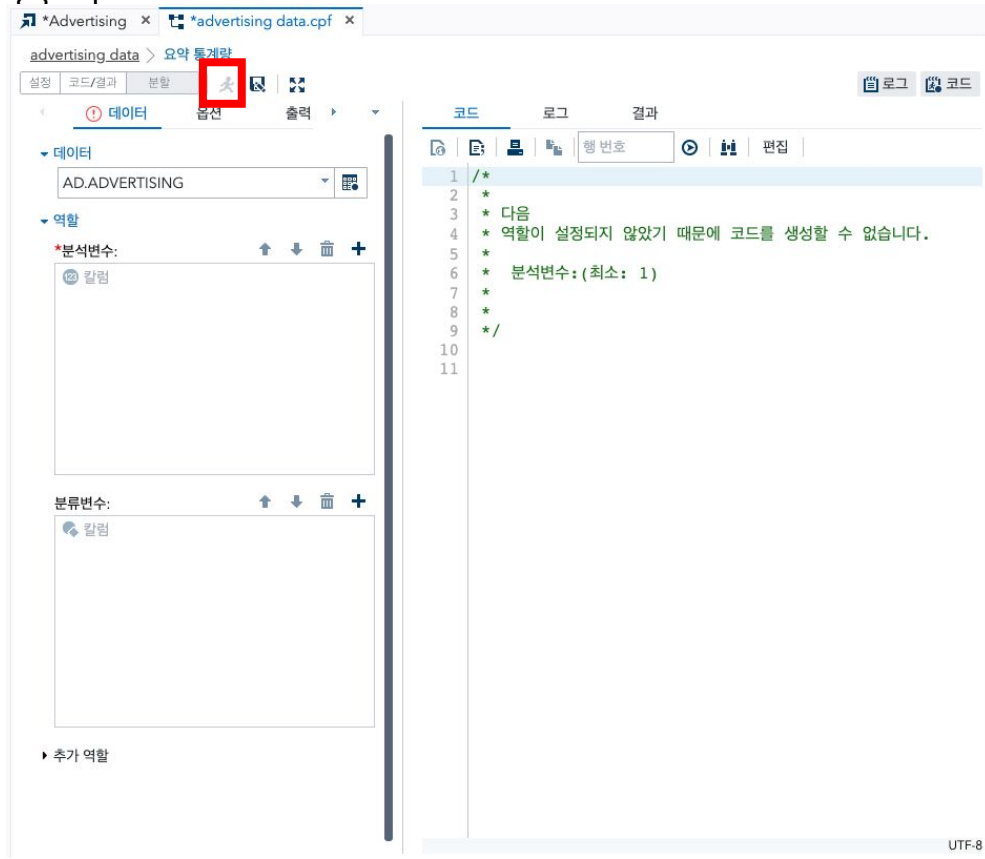
작업 및 유틸리티 > 작업 > 통계량 > 요약 통계량 > 데이터를 요약 통계량 노드에 연결 > 요약 통계량 더블클릭





# I 요약 통계량 인터페이스

대부분의 노드는 유사한 구조의 인터페이스를 가짐  
별표(\*)가 있는 항목은 필수적으로 지정해야하는 항목

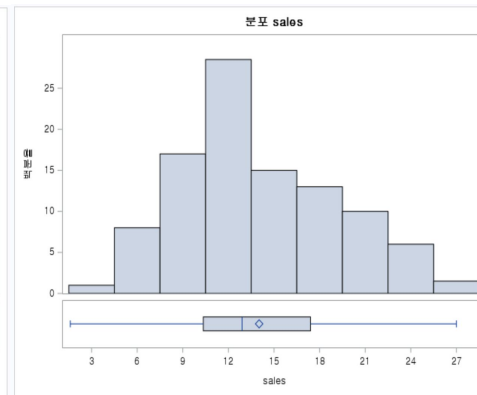
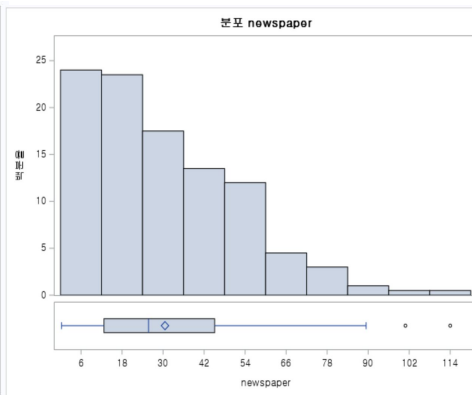
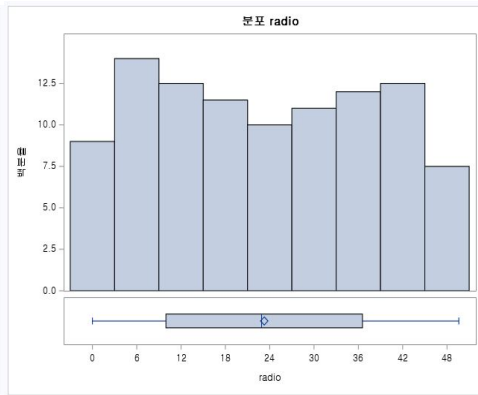
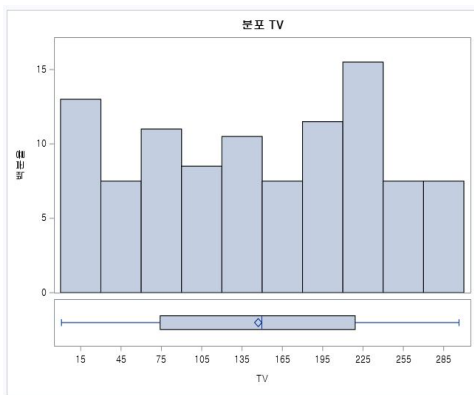


## I 요약 통계량 결과 화면

요약 통계량 결과 화면에서 전반적인 데이터에 대해  
점검

본 데이터에는 결측값이 존재하지 않음  
결측값이 있을 경우 제거하거나 다른 값으로  
대체해줘야함

변수	평균	표준편차	최솟값	최댓값	N	결측값 수	왜도	첨도
TV	147.0425000	85.8542363	0.7000000	296.4000000	200	0	-0.0698534	-1.2264948
radio	23.2640000	14.8468092	0	49.6000000	200	0	0.0941746	-1.2604014
newspaper	30.5540000	21.7786208	0.3000000	114.0000000	200	0	0.8947204	0.6495019
sales	14.0225000	5.2174566	1.6000000	27.0000000	200	0	0.4075714	-0.4088692





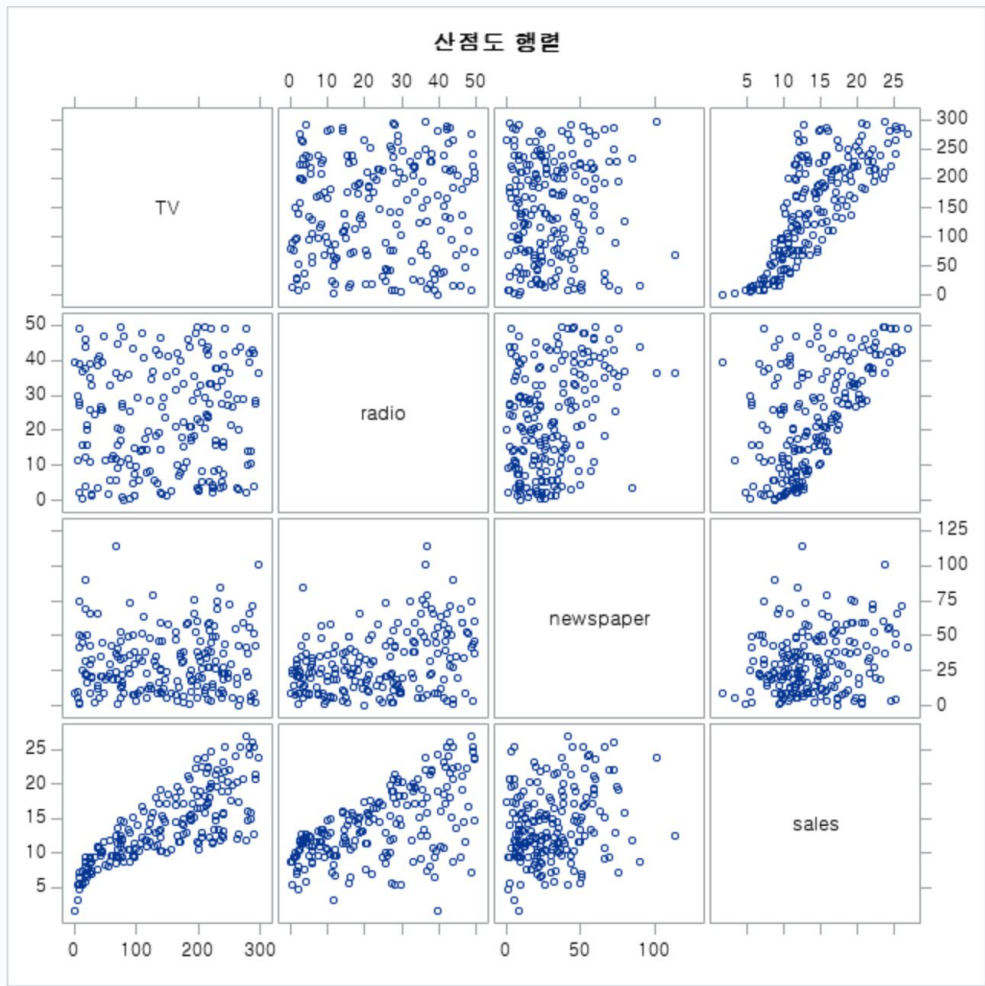
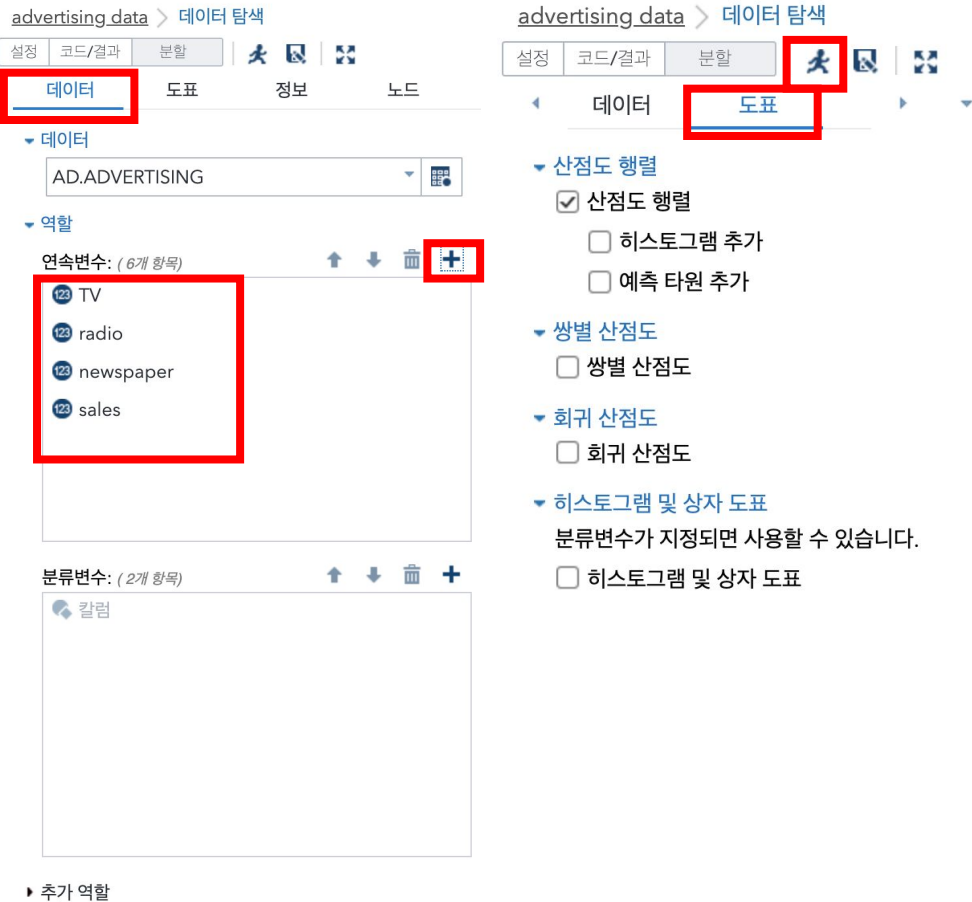
# I 데이터 탐색

작업 및 유틸리티 > 작업 > 통계량 > 데이터 탐색 > 데이터를 데이터 탐색 노드에 연결 > 데이터 탐색 더블클릭

The screenshot displays the FAST CAMPUS ONLINE interface. On the left, a sidebar menu shows the navigation path: '작업 및 유틸리티' (Job and Utility) is expanded, followed by '작업' (Job), '통계량' (Statistics), and '데이터 탐색' (Data Exploration). The '데이터 탐색' node is highlighted with a red box. On the right, the main workspace shows a flowchart with two nodes: 'ADVERTISING' and '데이터 탐색' (Data Exploration). The '데이터 탐색' node is also highlighted with a red box. The workspace includes tabs for '실행' (Execution), '결과' (Result), and '속성' (Property), and a toolbar with various icons for file operations and execution.

# I 데이터 탐색

데이터 &gt; 연속변수 선택 &gt; 도표 &gt; 산점도 행렬 &gt;



# I 분포분석

작업 및 유틸리티 > 작업 > 통계량 > 분포분석 > 데이터를 분포분석 노드에 연결 > 분포분석 더블클릭

The screenshot displays the FAST CAMPUS ONLINE interface. On the left, a sidebar menu shows the navigation path: '작업 및 유틸리티' (highlighted with a red box) > '작업' > '통계량' > '분포분석' (also highlighted with a red box). The main workspace shows a workflow diagram. It starts with a node labeled 'ADVERTISING'. Three arrows lead from this node to three subsequent nodes: '요약 통계량' (Summary Statistics), '데이터 탐색' (Data Exploration), and '분포분석' (Distribution Analysis). The '분포분석' node is highlighted with a red box. Each node contains a brief description of its function, and the '분포분석' node has a red circle icon at the bottom right.

# I 분포분석

데이터 > 연속변수 선택 > 도표 > 산점도 행렬 >

RUN

advertising\_data > 분포분석

설정 코드/결과 분할 **도표** 정보 노트

데이터 **옵션** 정보 노트

▼ 데이터 탐색

☒ 히스토그램

분류변수: (2개 항목)

칼럼

☐ 정규 곡선 추가

☐ 커널 밀도함수 추정값 추가

☐ 인셋 통계량 추가

▼ 정규성 확인

☐ 히스토그램 및 적합도 검정

☐ 정규 확률 도표

☒ 정규 Q-Q 도표

☐ 인셋 통계량 추가

▶ 적합 분포

