6.6 탐색 파일럿 실행 4단계 – 주제영역 3

이상 운전 패턴 스마트카 정보

과제 II

다음 페이지 부터 내용들을 실습해서 주제영역 3

“이상 운전 패턴 스마트카 정보” 마트를 생성 하세요.

주제 영역 3. 이상 운전 패턴 스마트카 정보 - 워크플로 작성 _ ⌨ 실습

주제 영역 3의 워크플로는 2020년 03월 22일에 스마트카 운전자의 운행 기록을 분석해서 과속, 급제동, 급회전이 빈번한 차량들을 스코어링한 마트 데이터를 생성한다. 과속과 급제동의 경우, 당일(22일)의 차량별로 가속 페달과 브레이크 페달의 평균값을 구하고, 관련 표준편차 값은 과거의 모든 데이터를 대상으로 산출해서 과속/급제동 표준값이 각각 "2" 이상인 차량의 경우만 "비정상(abnormal)"인 차량으로 판단했다. 급회전의 경우 당일(22일) 기준으로 Left/Right 회전각 "2-3" 단계를 "1000"번 이상 했을 경우 급회전이 빈번한 "비정상(abnormal)" 차량으로 지정했다. 워크플로의 하이브 작업에 사용되는 하이브 QL은 C://예제소스/bigdata2nd-master/CH06/HiveQL/의 경로에서 제공되므로 필요 시 해당 파일을 열어 복사/붙여넣기 하면 된다.

01. 휴의 좌측 드롭박스 메뉴에서 [문서]를 선택해 [내 문서]에 생성해 놓은 주제 영역 3의 작업 디렉터리로 이동한다.

- 휴 내 문서: /workflow/hive_script/subject3

02. 주제 영역 3에서는 사용할 하이브 스크립트 파일을 두 개 작성한다. 먼저 내 문서의 /workflow/hive_script/subject3으로 이동해서 [새 문서] → [Hive 쿼리]를 차례로 선택한다.

03. 하이브 에디트 창이 나타나고 운전자의 이상 운행 패턴을 관리하기 위한 하이브 테이블 스크립트를 작성하고 [저장] 버튼을 클릭한다. 파일명은 “create_table_managed_smartcar_symptom_info.hql”로 지정한다.

그림 6.102 주제 영역 3의 Managed 테이블을 생성하는 하이브 쿼리

04. 계속해서 내 문서의 /workflow/hive_script/subject3 위치에 두 번째 하이브 스크립트 파일을 만들어 본다. “subject3” 디렉터리에서 [새 문서] → [Hive 쿼리]를 선택한다.

05. 하이브 에디트 창이 나타나고 운전자의 이상 운행 패턴을 Select/Insert하는 하이브 쿼리 스크립트를 작성하고 [저장] 버튼을 클릭한다. 파일명은 “insert_table_managed_smartcar_symptom_info.hql”로 지정한다.

Hive Add a nam... Add a des...

```
insert into table Managed_SmartCar_Symptom_Info
select
        t1.car_number,
        t1.speed_p_avg_by_carnum,
        case
          when (abs((t1.speed_p_avg_by_carnum - t3.speed_p_avg) / t4.speed_p_std)) > 2
            then 'abnormal'
          else   'normal'
        end
        as speed_p_symptom_score,
        t1.break_p_avg_by_carnum,
        case
          when (abs((t1.break_p_avg_by_carnum - t3.break_p_avg) / t4.break_p_std)) > 2
            then 'abnormal'
          else   'normal'
        end
        as break_p_symptom_score,
        t2.steer_a_count,
        case
          when (t2.steer_a_count)  >   2000
            then 'abnormal'
          else   'normal'
        end
        as steer_p_symptom_score,
        t1.biz_date
```

가속 페달

브레이크 페달

운전대

(계속)

```
from
        (select car_number,
                biz_date,
                avg(speed_pedal) as speed_p_avg_by_carnum,
                avg(break_pedal) as break_p_avg_by_carnum
         from managed_smartcar_drive_info
         where biz_date = '${working_day}'
         group by car_number, biz_date) t1
join
        (select car_number,
                count(*) as steer_a_count
         from managed_smartcar_drive_info
         where steer_angle in ('L2','L3','R2','R3') and
               biz_date =  '${working_day}'
         group by car_number) t2
on
        t1.car_number = t2.car_number ,
        (select avg(speed_pedal) as speed_p_avg,
                avg(break_pedal) as break_p_avg
         from managed_smartcar_drive_info ) t3,
        (select stddev_pop(s.speed_p_avg_by_carnum) as speed_p_std,
                stddev_pop(s.break_p_avg_by_carnum) as break_p_std
         from (select car_number,
                      avg(speed_pedal) as speed_p_avg_by_carnum,
                      avg(break_pedal) as break_p_avg_by_carnum
               from managed_smartcar_drive_info
               group by car_number) s) t4
```

그림 6.103 주제 영역 3의 Managed 테이블에 데이터를 생성하는 하이브 쿼리

그림 6.103의 하이브 쿼리는 한 번의 실행으로 "스마트카 운전자 운행 정보(Managed_SmartCar_Drive_Info)"로부터 차량 번호별 스피드 페달, 운전대, 브레이크 페달의 데이터 분석 결과를 Managed_SmartCar_Symptom_Info 테이블에 저장한다. 쿼리를 자세히 살펴보면 전체 평균과 표준편차 값을 구하고, 당일(2020년 03월 22일)에 차량별 편차를 구해 이상 차량임을 판단하고 있는데, 이러한 처리 과정을 데이터의 피처 엔지니어링(Feature

Engineering)이라고 하며, 기존의 변수를 가공해 새로운 변수와 정보를 추가하는 과정에 해당한다. 참고로 해당 쿼리가 실행될 때 Job Browser를 모니터링해 보면 7개의 잡과 10여 개 이상의 맵리듀스가 실행되는 것을 확인할 수 있다. 총 수행 시간도 필자의 파일럿 환경에서 10여 분 정도 소요되는 다소 무거운 하이브 쿼리다.

06. 이제 워크플로를 만든다. 휴 상단의 쿼리 콤보박스의 [스케줄러] → [Workflow]를 선택해 우지 편집기로 이동한다.

07. 첫 번째 작업으로 워크플로의 작업 툴박스에서 “Hive 쿼리” 작업을 워크플로의 첫 번째 작업 노드에 드래그 앤드 드롭한다.

08. 사용할 Hive 스크립트 파일을 선택한다. 앞 단계에서 내 문서의 /workflow/hive_script/subject3에 만들어 둔 create_table_managed_ smartcar_symptom_info.hql을 선택한 후 [추가] 버튼을 클릭한다.

09. 두 번째 작업을 위해 워크플로의 작업 툴박스에서 “Hive 쿼리” 작업을 워크플로의 두 번째 작업 노드에 드래그 앤드 드롭한다.

10. 사용할 Hive 스크립트 파일을 선택한다. 앞 단계에서 만든 내 문서의 /workflow/hive_script/subject3에 insert_table_managed_smartcar_symptom_info.hql을 선택한 후 [추가] 버튼을 클릭한다.

11. [매개변수+]를 누르고 working_day의 매개변수에 우지의 예약 스케줄러에서 정의할 "${today}" 매개변수를 할당한다.

   - working_day=${today}　※ 즉시 실행시 - 스마트카 시뮬레이션 날짜(교재기준: 20200322)로 설정
     ex) working_day=20200322

12. 워크플로의 이름을 작성한다. 워크플로 상단의 [My Workflow]를 클릭하고, "Subject 3 – Workflow"로 변경한 후 [확인] 버튼을 누른다.

13. 워크플로 작성을 완료한다. 우측 상단의 [저장] 버튼을 누른다.

14. 이제 작성한 워크플로를 작동하기 위한 예약 작업을 생성한다. 상단의 쿼리 콤보박스에서 [스케줄러] → [예약]을 차례로 선택한다.

15. 먼저 예약 작업 이름을 입력한다. 상단의 [My Scheduler]를 클릭하고 "Subject 3 – 예약"으로 입력한다.

16. 예약 작업이 사용할 워크플로를 선택한다. "예정된 Workflow는 무엇입니까?"라는 메시지 하단에 있는 "Workflow 선택…"을 클릭해 앞서 만든 주제 영역 3의 워크플로인 "Subject 3 – Workflow"를 선택한다.

17. 예약 작업 워크플로를 실행시키기 위한 스케줄 값을 입력한다.

- **실행 간격:** 매일, 03시
- **시작 일자:** 2020년 03월 23일, 00시 00분
- **종료 일자:** 2020년 12월 31일, 23시 59분
- **시간대:** Asia/Seoul

18. 워크플로에서 사용할 매개변수인 today 값을 예약 작업의 매개변수로 정의한다.
    앞서 워크플로의 하이브 작업에서는 매개변수를 "working_day=${today}"로 등록했다. today의 값을 Coordinator의 내장 함수를 통해 설정한다.

```
${coord:formatTime(coord:dateTzOffset(coord:nominalTime(), "Asia/Seoul"), 'yyyyMMdd')}
```

19. 우지의 예약 작업 설정이 모두 끝났다. [저장] 버튼을 클릭해 작성을 완료한다.

20. 작성이 완료된 예약 작업을 우측 상단의 [제출] 버튼을 클릭해 실행한다.

21. 제출된 예약 작업 상태를 확인해 본다. 좌측 상단의 드롭박스 메뉴에서 [Job] → [일정]을 차례로 선택한다. 앞서 등록한 "Subject 3 – 예약"이 "Running" 상태로, 매일 새벽 03시가 되면 등록된 워크플로를 작동시키게 된다. 새벽 3시까지 기다릴 수 없으니 앞서 주제 영역 1에서 설명한 "워크플로 즉시 실행해 보기"를 참고해 곧바로 실행해 본다.

22. “Subject 3 – Workflow”가 정상적으로 작동했는지 확인한다. 상단의 쿼리 콤보박스에서 [편집기] → [Hive]를 선택해 하이브 에디터에서 그림 6.105와 같이 하이브 조회 쿼리를 작성해 실행한다. 참고로 “biz_date=20200322”의 날짜는 독자들의 파일럿 환경의 biz_date 날짜로 맞춰야 한다.

Hive insert_table_m... Add a descripti...

1m, 8s Database default 유형 text

```
SELECT
    car_number,
    cast(speed_p_avg as int),
    speed_p_symptom,
    cast(break_p_avg as float),
    break_p_symptom,
    cast(steer_a_cnt as int),
    steer_p_symptom,
    biz_date
FROM managed_smartcar_symptom_info
where biz_date = '20200322'
```

쿼리 기록 저장된 쿼리 결과 (94)

	car_number	speed_p_avg	speed_p_symptom	break_p_avg	break_p_symptom	steer_a_cnt	steer_p_symptom
1	A0004	1	normal	0.30251548	normal	1559	normal
2	A0097	1	normal	0.3166468	normal	1568	normal
3	B0006	1	normal	0.30939516	normal	1552	normal
4	B0043	1	normal	0.30955213	normal	1558	normal
5	B0051	1	normal	0.3005024	normal	1544	normal
6	B0065	1	normal	0.28641355	normal	1449	normal
7	B0070	2	abnormal	0.20799153	abnormal	2959	abnormal

그림 6.105 주제 영역 3 워크플로의 실행 결과 확인

23. “비정상 스마트카 운행” 데이터를 차트로 재구성해 보면 좀 더 직관적으로 데이터를 탐색할 수 있다. 먼저 그림 6.106의 하이브 쿼리 실행 결과에서 [차트] 버튼을 선택한다.

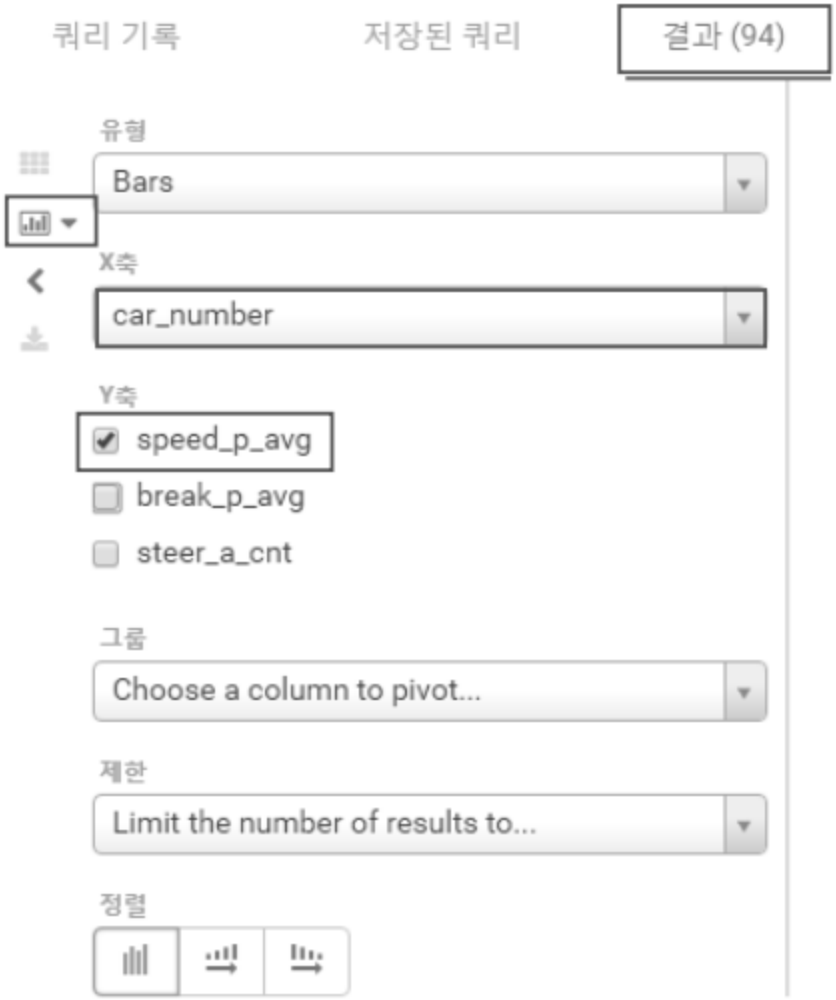

그림 6.106 주제 영역 3 워크플로의 실행 결과 확인 – 차트 선택

차트의 종류로 “Bars(막대)”를 선택한다. 첫 번째로 가속 페달의 비정상 패턴 차량을 조회해서 과속/난폭 운전 가능성이 예상되는 차량들을 찾아보자.

이를 위해 X축에서는 “car_number”를 선택하고, Y축에서는 “speed_p_avg”를 선택한다.

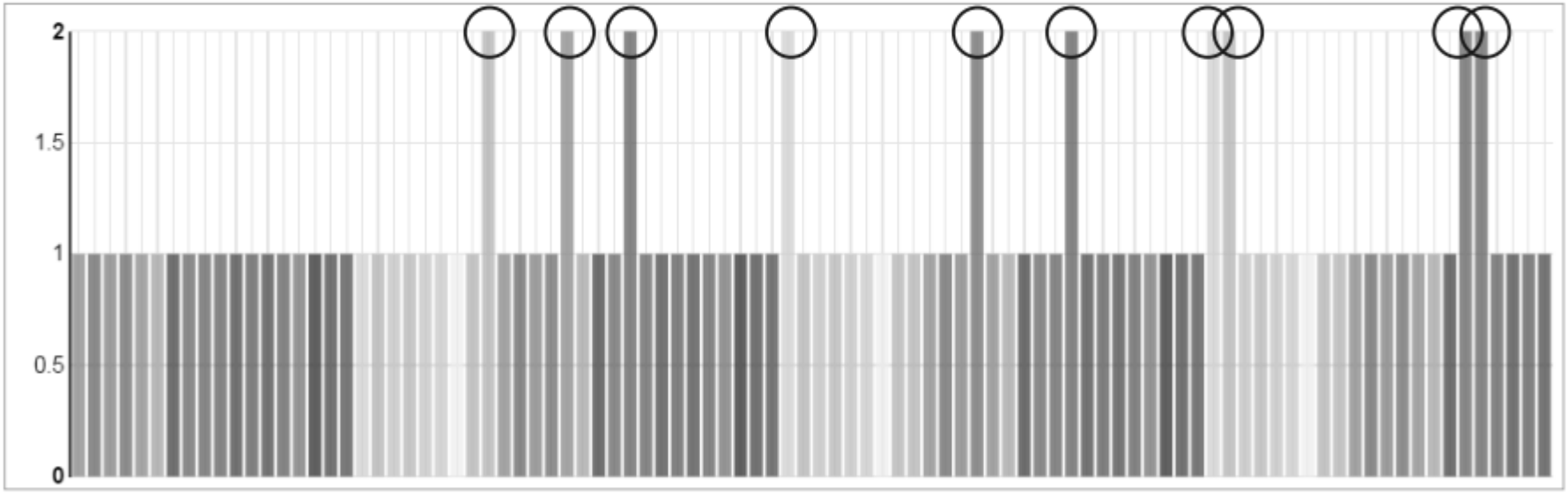

그림 6.107 주제 영역 3 워크플로의 실행 결과 확인 – 차트 보기 1

조회 결과를 보면 차량별 가속 페달의 편차가 매우 높은 스마트카 차량 10대가 발견됐다.

두 번째로 브레이크 페달에서 비정상 패턴을 보이는 차량을 조회해서 급정지/난폭 운전의 가능성이 예상되는 차량들을 찾아보자.

이를 위해 X축에서는 “car_number”를 유지하고, Y축에서는 “speed_p_avg” 선택을 해제하고 “break_p_avg”를 선택한다.

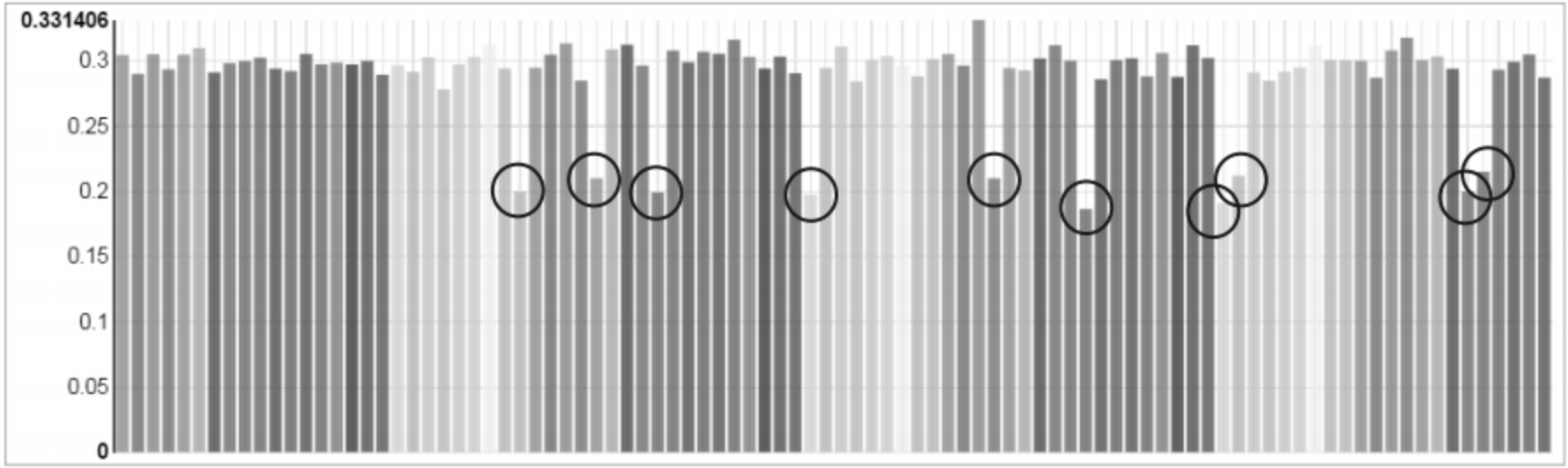

그림 6.108 주제 영역 3 워크플로의 실행 결과 확인 – 차트 보기 2

조회 결과를 보면 차량별 브레이크 페달의 편차가 지나치게 낮은 스마트카 차량 10대가 발견됐다.

세 번째로 운전대의 비정상 패턴을 보이는 차량을 조회해서 급회전/난폭 운전의 가능성이 예상되는 차량들을 찾아 보자.

이를 위해 X축에서는 “car_number”를 유지하고, Y축에서는 “break_p_avg” 선택을 해제하고 “steer_a_cnt”를 선택한다.

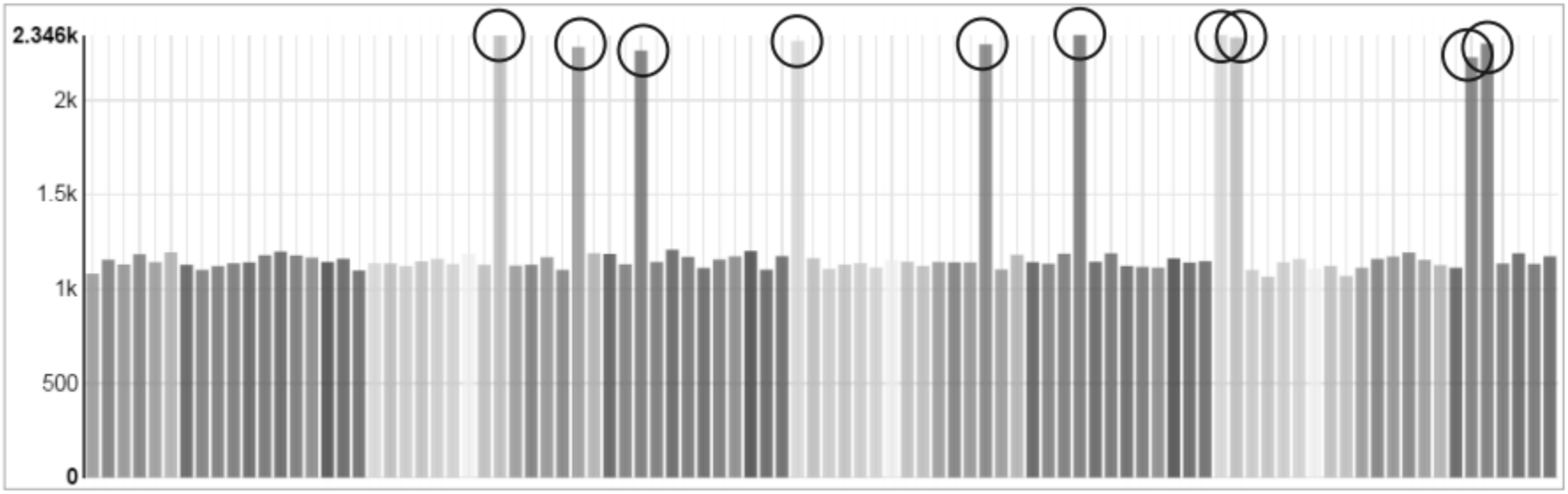

그림 6.109 주제 영역 3 워크플로의 실행 결과 확인 – 차트 보기 3

조회된 결과를 보면 차량별 운전대 회전각이 평균보다 지나치게 높은 스마트카 차량 10대가 발견됐다.

위 3개의 차트를 분석해 보면 이상 운행 패턴을 보이는 운전자는 가속 페달, 브레이크 페달, 운전대 사용 패턴도 모두 비정상인 것으로 파악됐다. 결국 3개의 변수(가속 페달, 브레이크 페달, 운전대)들은 서로 연관성이 매우 높다는 사실을 알 수 있다.

주제 영역 3에서는 하이브의 단순 기술적 통계량으로 이상 징후 차량들을 탐색했다. 여기서 스코어링된 차량 중 특정값에 비정상 판정을 받은 차량의 경우 속도, 브레이크, 회전에 있어서도 다른 차량보다 수치가 지나치게 크거나/낮게 나타난다는 것을 확인했다. 하이브를 잘 이용하면 빅데이터에 적재돼 있는 대용량 데이터도 쉽고 빠르게 탐색적 분석을 할 수 있다.

Tip _ 빅데이터 분석을 위한 탐색 및 전처리 작업

일반적으로 빅데이터 인사이트는 탐색 단계에서 80% 이상이 발견되고, 나머지 20%는 분석 단계에서 검증하며 얻게 된다. 특히 탐색 단계에서는 데이터의 전처리 작업의 비중이 매우 높은데 일련의 과정들을 그림 6.104로 정의할 수 있다. 크게 4개의 단계가 있으며, 사용하는 분석 기법과 알고리즘, 분석 환경에 맞게 선택적으로 수행하게 된다. 파일럿 프로젝트에서는 그림 6.104의 모든 전처리 작업을 다루지는 않는다. 빅데이터 모델러, 데이터 엔지니어, 분석가에 관심 있는 독자라면 향후 이 분야의 이론과 기술들을 좀 더 집중적으로 공부해 보길 바란다.

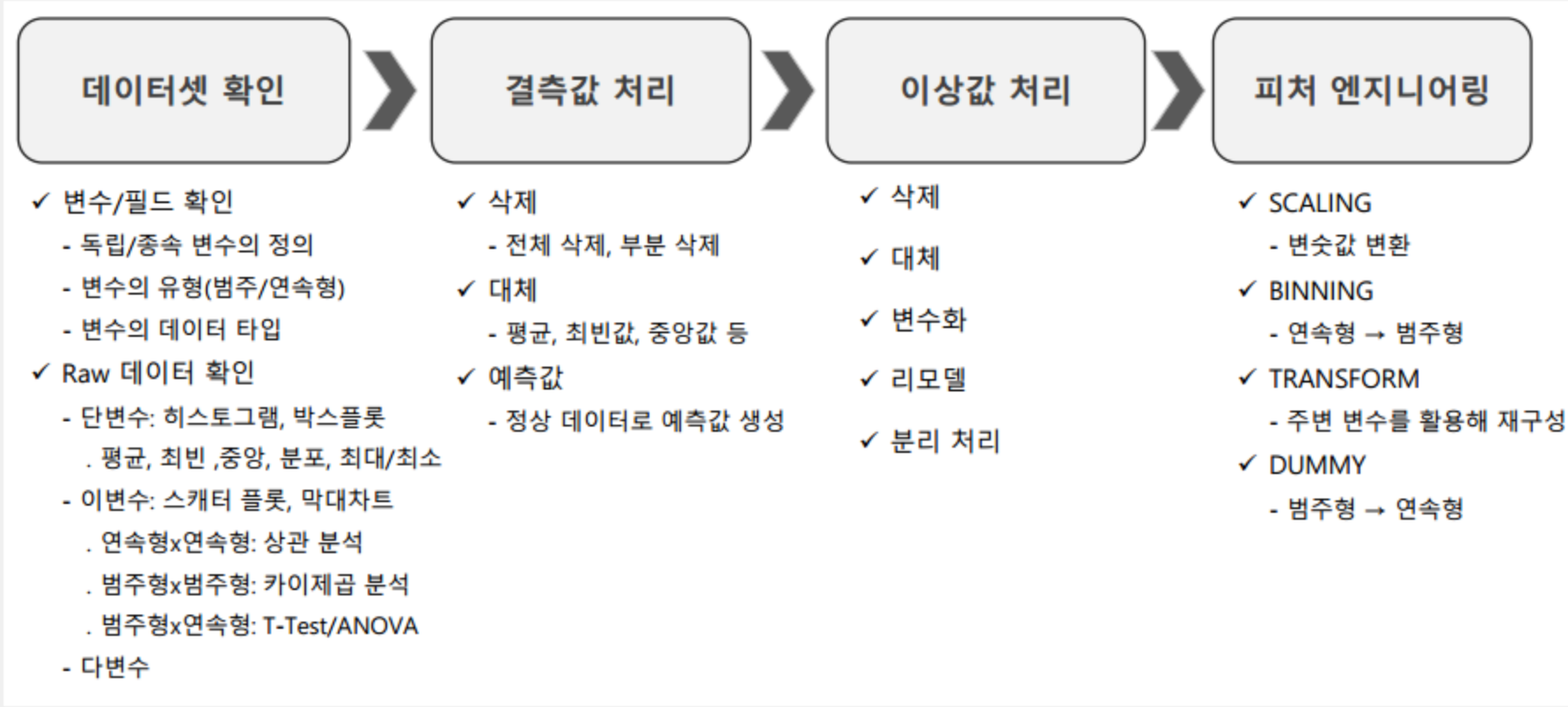

그림 6.104 빅데이터 분석을 위한 탐색 및 전처리 작업