

1-Cycle 데이터셋 정제/가공 절차 및 결과 안내 가이드

개발(참여)기관	㈜에티포스	최신 배포일	2022.10.25	
개발/작성자	강성현	승인자	참여(개발)기관	주관기업
			오정찬	
과제명	11번. 자율주행차 센서 분석 데이터 – 인공지능 학습용 데이터 구축 사업			
개발명	V2X 메시지 데이터 정제 및 가공			
개발이력	일자	2022.10.24	Version.	v.08
	일자		Version.	
	일자		Version.	
	일자		Version.	
	일자		Version.	

[사업계획 및 데이터셋 개발작업 기준]

3.2.2 데이터 정제 도구

- 중복 제거
 - 중복 제거 프로그램을 직접 제작함
 - 동일 시간에 생성된 데이터 기준으로 데이터가 완전 같은 경우 하나의 데이터를 제거
- 데이터 자르기
 - 데이터 자르기 프로그램을 직접 제작함
 - PVD가 포함하는 다양한 데이터 중 위치/방향/속도 등 차량의 운행과 연관된 데이터만 남기고 다른 부분은 잘라내는 기능을 수행
- 비식별화
 - 비식별화 프로그램을 직접 제작함
 - 실제 운행 차량을 식별할 수 있는 차량의 고유 ID를 hash 등 방안을 이용하여 익명화된 ID로 치환하는 기능을 수행
- 시계열 데이터화
 - 변환 프로그램을 직접 제작함
 - 차량 ID, 시간별 데이터 정렬, 데이터 묶음, 데이터 간 차이 값 계산, 정규화를 통한 시계열 데이터 생성 및 CSV형식 저장을 수행

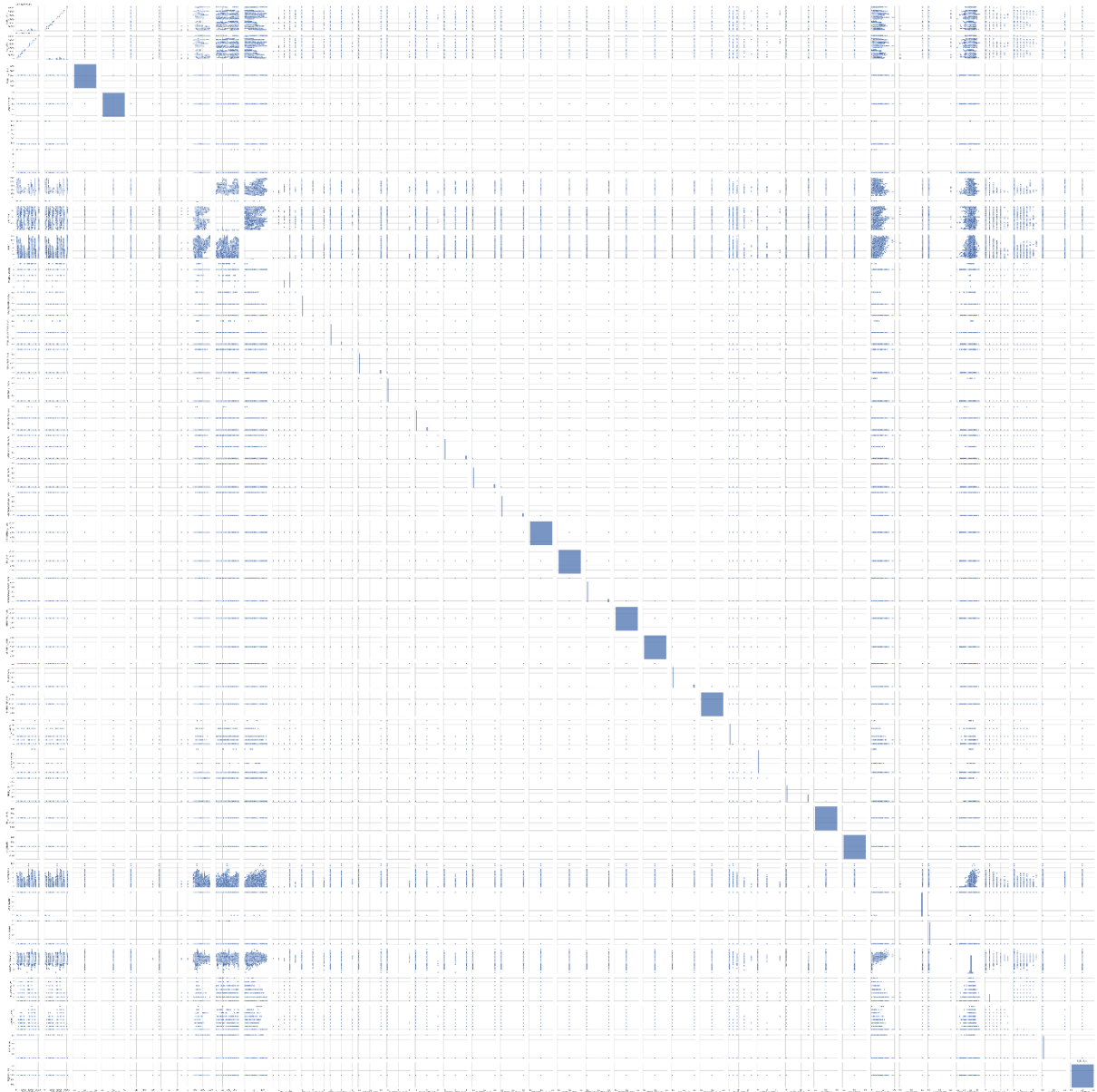
1. **중복 제거**

- 1) 수집된 데이터에서 GPS 상태에 따라 위치정보 미수집된 데이터는 존재하지 않았음.
- 2) 동일 시간에 생성된 데이터를 기준으로 데이터가 완전 같은 경우에 삭제라고 되어있어 모든 row를 기준으로 값이 똑같은 데이터가 있는지 검사한 결과, 존재하지 않았음.
- 3) 그러나 **ISSEUE_DATE**와 **OBU_ID**를 기준으로 중복되는 값을 검사한 결과, **RSU_ID**를 제외한 나머지 데이터가 모두 같은 중복 데이터가 존재함.
이는 동일한 메시지 데이터가 두 군데 이상의 RSU에서 중복 수신된 경우로 간주하여 중복데이터 중 두 번째에 나온 데이터부터 제거함.

2. **데이터 검토 및 EDA 분석**

Raw 데이터로부터 주요 객체 정보 변수와 목표로 하는 차량 동작 분류를 한 후보 변수로, 세종시의 총 42개 컬럼 데이터를 탐색하고 그중 십 여개의 기대되는 관련 주요 변수들로 Plotting 등 시각화 및 상관관계 분석을 수행하였으나, 당연하게도 차량 속도값('SPEED')과 '연료분사 노즐 개구율('THROTTLEPOS') 두 변수간 약 0.56의 상관계수 값을 갖는 유의한 상관관계를 보인 것 외에는 변수간 큰 상관성을 보이는 항목은 없는 것으로 보임.

1) 먼저 주어진, 광주시 V2X 메시지 샘플 데이터(10,000 rows)로 1차 검토를 위한 21개 변수들을 Scatter Plotting을 수행.



2) 입수된, 9월 세종시 데이터 중, 가장 많은 메시지 및 데이터 크기를 갖는 22년 09월 26일 데이터를 바탕으로 약 13~14개의 변수들로 Plotting 및 상관계수 산출과 Heatmap 시각화를 수행하여 변수간 상관성을 확인한 결과, 상기 광주 데이터처럼 변수간에 영향도가 큰 변수는 서로를 개구울 외에는 발견하기 힘들.

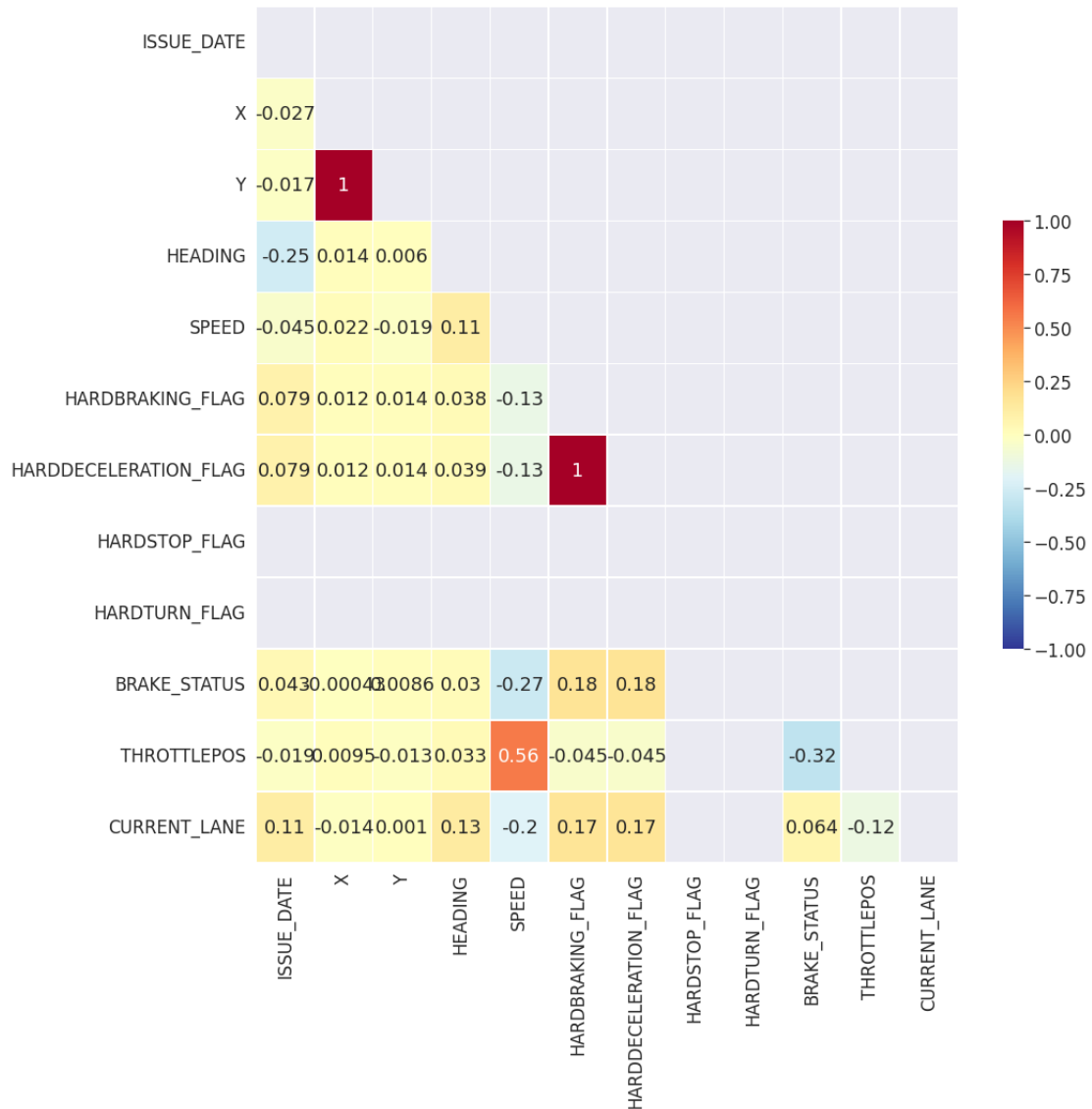
[참고1] 제공된 세종시의 09월01일부터 09월 30일까지 한달치 데이터를 모두 병합하여 Plotting을 수행하였으나 병합(merge)데이터의 row 수는 약 116,450여건으로 구글 Colab 환경에서는 RAM 부족으로 인한 세션 종료가 됨.

[참고2] 아래 이미지는 세종시 9월 26일자 데이터를 기준으로 Plotting, 상관관계수 산출 및 Heatmap 형태로 시각화한 이미지 임



< 13Var_Scattering >

```
> 13 Variables : Index(['ISSUE_DATE', 'OBU_ID', 'X', 'Y', 'HEADING', 'SPEED',
                        'HARDBRAKING_FLAG', 'HARDDECELERATION_FLAG', 'HARDSTOP_FLAG',
                        'HARDTURN_FLAG', 'BRAKE_STATUS', 'THROTTLEPOS', 'CURRENT_LANE'],
```



< 13Var_Corr. HeatMap >

3. **데이터 자르기**

1) 주요 목표 변수 및 관련이 있을 수 있는 변수열(29개 변수열)만 남기고 1차 제거 후, 재 검토 및 분석 목표에 필요한 핵심 변수로 객체 정보 및 차량 운행과 연관된 컬럼으로 최종 선정(15개 변수열)하고 다른 컬럼 제거.

> dataframe name : main_df

```
> 15개 변수명 : Index(['ISSUE_DATE', 'OBU_ID', 'VEHICLE_CLASS',
'VEHICLE_TYPE', 'X', 'Y', 'HEADING', 'SPEED', 'HAZARDLIGHTS_FLAG',
'HARDBRAKING_FLAG', 'LIGHTSCHANGE_FLAG', 'HARDDECELERATION_FLAG',
'UTURN_FLAG', 'LIGHTS_STATUS', 'CURRENT_LANE'])
```

4. **비식별화**

1) 차량의 고유 ID(OBU_ID)를 기준으로 비식별화를 진행하기 위해 groupby 메서드를 통하여 새로운 dataframe 생성.

> dataframe name : de_ident_df

2) 비식별화 ID는 8자리(String Type)로 생성하였으며 'random.choices' 메서드를 통해 알파벳 대문자와 숫자를 랜덤으로 혼합하였음.

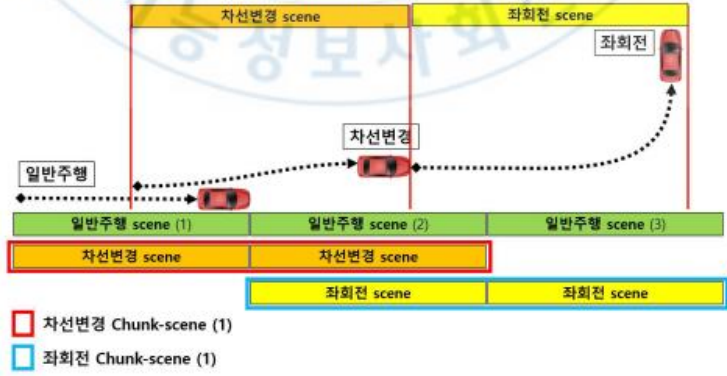
3) 혼합된 비식별화 ID는 'random_id' 이름의 list에 저장.

4) 원천 데이터프레임에 있는 OBU_ID를 모두 바꾸기 위해서는 기존 OBU_ID와 새로 생성한 비식별화 ID를 dictionary로 mapping함.

5) 이를 바탕으로 'replace' 메서드를 활용하여 원천 데이터프레임에 있는 OBU_ID를 모두 비식별화 함.

5. **시계열 데이터화**

나) 정제 항목

구분	내용	비고
정제데이터 구성	<ul style="list-style-type: none"> - 수집 데이터에서 GPS 상태에 따라 위치정보의 미수집, 완전히 동일한 메시지를 2번이상 수신된 경우 등이 발생된 경우 해당 데이터는 버림 - Scene을 구성할 때 10초간 2개 이상의 메시지가 누락/손실 되는 경우 Scene을 생성하지 않음 ※V2X 통신의 요구조건인 패킷에러율 10%을 기준으로 함 - 수집 및 정제 과정에서 오류로 인한 메시지 손실 및 시계열 데이터 내 메시지 누락이 1건 발생 시 interpolation을 통해 메시지 보간 	
정제기준	<ul style="list-style-type: none"> - SAE J2735 V2X 메시지 표준에 따라 수집된 데이터 정제 - 동일 객체에 대하여 10초간의 메시지를 묶어 1개의 Scene으로 정의 ※ SAE 표준문서에서 1Scene을 10초로 권장하고 있음 	
Event-Scene 정의	<ul style="list-style-type: none"> - 클래스에 해당하는 사건의 종료 시점을 기준으로 20초 이전까지의 시간을 이벤트로 정의 - 이벤트가 포함된 Scene들을 하나로 묶어 Event-Scene으로 정의 	

차량을 기준으로 'groupby'를 진행하였고, 1초 단위로 데이터프레임의 'resample' 수행하였으며,

이 과정에서 누락된 시간의 데이터는 NaN 값으로 대체 후, null값이 1개 초과인 경우를 'scene_df'

데이터프레임에 새로 저장함.

1) 수집 및 정제 과정에서 오류로 인한 메시지 손실 및 시계열 데이터 내 메시지 누락이 1건 발생시 interpolation을 통해 메시지 보간 수행.

> interpolation() 함수를 통해 이전 값과 이후 값의 평균치 값을 지정

2) 10초간 메시지를 구성하는데 있어 2개 이상의 메시지가 누락/손실되는 경우 Scene을 생성하지 않음.

> scene_df에서 null값이 2개 이상인 경우에 바로 drop을 하도록 구성.

3) Event 단위로 보기 위해 20초씩 데이터 차이를 본 후, 이를 'car_1_diff' 및 'car_diff' 데이터프레임에 저장. (최종 데이터프레임은 'car_diff')

****ISSUE 사항****

1) 연속된 11개의 데이터를 선정한다는 의미는, 시계열 데이터에 있어 중복됨을 의미함.

(예: 1초~10초, 10초~20초 > 10초 포인트가 중복됨)

> 이렇게 진행 시 시계열이 깨지게 되므로 10개의 메시지를 기준으로 'scene' 단위 선정

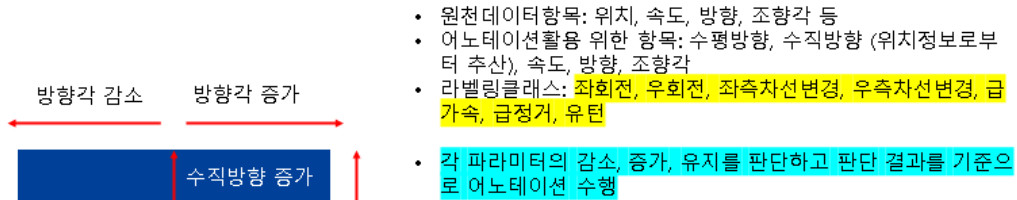
2) 위도와 경도, 속도 차이를 정규화한다면 후속 가공(7개 차량 행동 클래스 분류 로직)이 진행이 어려워질 뿐만 아니라 역추적 또한 어려워, 실제 해당 값으로 이벤트 구성 및 분류 로직 개발

> 차량 동작 클래스 분류를 위한 데이터 가공 진행 후, 필요시 정규화 진행 or

GPS 데이터 삭제 가능

6. **차량 동작 클래스화 가공(자동 라벨링)**

* 가공 방안 *



문의	좌회전	우회전	좌측차선변경	우측차선변경	급가속	급정거	유턴
수평방향	감소	증가	감소	증가	-	-	감소후 불거
수직방향	증가후 불거	증가후 불거	증가	증가	-	-	불거후 감소
속도	-	-	-	-	감소	증가	-
방향	감소후 불거	증가후 불거	감소후 불거	증가후 불거	-	-	감소후 불거
조향각	감소후 증가	증가후 감소	감소	증가	-	-	감소후 증가

*각 파라미터의 감소, 증가, 유지를 판단하기 위한 기준값은 데이터 수집/정제 후 테스트를 통해 결정

* 가공 항목 *

○ 라벨링 작업 대상

구분	내용	비고																
Scene 단위로 정제되어 구성된 V2X 메시지 데이터	<ul style="list-style-type: none"> - 정제된 원천 데이터에 Python으로 개발된 V2X 메시지 어노테이션 도구를 이용하여 좌/우회전, 차선변경, 급가속/급정거, 유턴, 돌발상황 등으로 정의된 클래스를 라벨링 - 자동화 과정을 거쳐 라벨링 된 이벤트 발생 메시지에 대해 End-Trigger 방식을 이용하여 20초간의 이벤트가 포함된 Scene들을 하나로 묶어 Event-Scene으로 정의 - 이렇게 정의된 Event-Scene에 대하여 작업자가 V2X 메시지 시각화 도구를 활용하여 차량의 실제 이벤트를 클래스로 설정 																	
프로그램을 이용한 자동 라벨링 기준	<table border="1"> <thead> <tr> <th>클래스명</th><th>기준</th><th>비고</th></tr> </thead> <tbody> <tr> <td>Turn(Right/Left)</td><td>Currentlane 변경 Steering±20~40° Heading±20~40°</td><td rowspan="6"> Currentlane : 현재차선 Steering : 바퀴의 각도 Heading : 차량의 방향 Speed : 차량의 속도 BrakeSystemStatus : 브레이크 상태 </td></tr> <tr> <td>Change(Right/Left)</td><td>Steering±40~50° Heading±80~100°</td></tr> <tr> <td>Speed(Acceleration)</td><td>Currentlane 변경없음 Heading/Steering±0~5° Speed+15km/h</td></tr> <tr> <td>Speed(Hardbrakes)</td><td>Currentlane 변경없음 Heading/Steering±0~5° Speed-15km/h BrakeSystemStatus On</td></tr> <tr> <td>Hazard(True)</td><td>BrakeSystemStatus Off Speed -15km/h</td></tr> <tr> <td>Turn(UTutn)</td><td>Currentlane 변경 Steering±45~60° Heading±165~190°</td></tr> </tbody> </table>	클래스명	기준	비고	Turn(Right/Left)	Currentlane 변경 Steering±20~40° Heading±20~40°	Currentlane : 현재차선 Steering : 바퀴의 각도 Heading : 차량의 방향 Speed : 차량의 속도 BrakeSystemStatus : 브레이크 상태	Change(Right/Left)	Steering±40~50° Heading±80~100°	Speed(Acceleration)	Currentlane 변경없음 Heading/Steering±0~5° Speed+15km/h	Speed(Hardbrakes)	Currentlane 변경없음 Heading/Steering±0~5° Speed-15km/h BrakeSystemStatus On	Hazard(True)	BrakeSystemStatus Off Speed -15km/h	Turn(UTutn)	Currentlane 변경 Steering±45~60° Heading±165~190°	
클래스명	기준	비고																
Turn(Right/Left)	Currentlane 변경 Steering±20~40° Heading±20~40°	Currentlane : 현재차선 Steering : 바퀴의 각도 Heading : 차량의 방향 Speed : 차량의 속도 BrakeSystemStatus : 브레이크 상태																
Change(Right/Left)	Steering±40~50° Heading±80~100°																	
Speed(Acceleration)	Currentlane 변경없음 Heading/Steering±0~5° Speed+15km/h																	
Speed(Hardbrakes)	Currentlane 변경없음 Heading/Steering±0~5° Speed-15km/h BrakeSystemStatus On																	
Hazard(True)	BrakeSystemStatus Off Speed -15km/h																	
Turn(UTutn)	Currentlane 변경 Steering±45~60° Heading±165~190°																	
시각화 도구 활용	<ul style="list-style-type: none"> - 메시지 데이터의 항목 중 차량의 위치 정보에 해당하는 위도, 경도 등이 수치로 표시되며, 작업자가 이 데이터만으로 차량의 실제 주행 경로를 파악하기 어려움 - 시각화 도구를 활용, 라벨링 대상 Scene의 실제 주행 경로 등을 파악하여 클래스 설정 - PVD 메시지에 GPS 오차가 존재하여 표시되는 위치와 실제 위치가 상이할 수 있으므로, 비식별화 된 CCTV등의 관련 동영상을 활용하여 실제 주행 경로 확인 																	

1) 가공항목은 앞서 붙여진 클래스명에 따라 해당 아이디어 기준에 맞게 데이터 조건식을 활용하여 데이터프레임 columns 생성.

> 기존 'car_diff' 데이터프레임의 컬럼 추가됨

※ 'Steering_Angle(조향각)'은 광주시 샘플데이터의 컬럼에는 반영되어 있으나 수집 정보가 없고, 현재 수집된 세종시 데이터에는 컬럼 등 데이터 자체가 없어 제외됨.

2) 추가된 컬럼명["~~"] 및 클래스 레이블은 분류 로직 및 변수 기준에 따라 아래와 같이 4개 컬럼에 7개 클래스가 통합 표시되도록 하였으며, 1-Cycle 대응 후 세종시 등 다른 날짜 및 지역 데이터에도 공통 적용 및 자동화 로직을 적용할 예정임

- ① car_diff["Turn"] : [right](#) / [left](#) / [Uturn](#) (좌/우회전 및 U턴)
- ② car_diff["Change"] : [right](#) / [left](#) (좌/우 차선변경)
- ③ car_diff["Speed"] : [Acceleration](#) / [Hardbrakes](#) (가속 / 급정거)
- ④ car_diff["Hazard"] : [True](#) / [False](#) (비상조건 True 여부)

[기타]

■ 해당 소스코드는 별첨 문서로 첨부되었으며, 향후 GitHub 링크로 공유될 예정

■ 생성된 정제/가공 데이터셋 및 소스코드 등은 아래 사업 산출물 저장소인 NAS의 해당 링크의 아래 디렉토리 폴더에 저장 및 관리 중

- 1-Cycle 결과 샘플 데이터셋

: /AI_2-050/00.구축데이터/1.노변기지국송수신V2X메시지데이터/3.가공데이터/3.1-Cycle

- 세종시 9월 데이터 정제/가공 세트

: /AI_2-050/00.구축데이터/1.노변기지국송수신V2X메시지데이터/3.가공데이터/2.세종9월/