제공 데이터

( train, test , bus\_bts 공통 사항)  
해당 데이터에는 버스카드를 통해 결제를 한 경우에 대한 정류소 승, 하차 데이터로 모든 승차정보의 경우는 기록이 되어있지만, 버스에서 하차를 할 때, 버스카드를 찍지 않는 경우, 해당 기록이 비어 있는 상태입니다. 따라서, 승차 인원수와 하차 인원수가 동일하지 않고 다소 차이가 있음을 미리 알려드립니다.

(train, test csv 공통사항)  
해당 버스정류장에 대한 각각의 위도, 경도가 제공이 되어있는 상태로 같은 정류장 이름이지만 위도와 경도가 서로 다른 경우가 존재합니다. 해당 경우는, 같은 정류장 이름을 가지고 있는 길 건너편의 정류장에 해당이 됩니다.

1. train.csv and test.csv

train.csv 의 경우, 2019년 9월 제주도의 각 날짜, 출근시간(6시~12시)의 버스 정류장별 승하차 인원, 퇴근시간(18시~20시)의 버스 정류장별 승차 인원이 기록되어 있습니다.

test.csv의 경우, 2019년 10월의 각 날짜, 출근시간(오전 6시~12시)의 버스 정류장별 승하차 인원이 기록되어 있습니다.

참가자들은 train.csv를 통해 퇴근시간 승차 인원을 예측하는 모델을 생성한 후 test.csv에서 실제 예측을 진행하면 됩니다.

|  |  |
| --- | --- |
| 변수명 | 데이터 설명 |
| id | 해당 데이터에서의 고유한 ID(train, test와의 중복은 없음) |
| date | 날짜 |
| bus\_route\_id | 노선ID |
| in\_out | 시내버스, 시외버스 구분 |
| station\_code | 해당 승하차 정류소의 ID |
| station\_name | 해당 승하차 정류소의 이름 |
| latitude | 해당 버스 정류장의 위도  (같은 정류장 이름이어도 버스의 진행 방향에 따라 다를 수 있음) |
| longitude | 해당 버스 정류장의 경도  (같은 정류장 이름이어도 버스의 진행 방향에 따라 다를 수 있음) |
| 6~7\_ride | 6:00:00부터 6:59:59까지 승차한 인원 수 |
| 7~8\_ride | 7:00:00부터 7:59:59까지 승차한 인원 수 |
| 8~9\_ride | 8:00:00부터 8:59:59까지 승차한 인원 수 |
| 9~10\_ride | 9:00:00부터 9:59:59까지 승차한 인원 수 |
| 10~11\_ride | 10:00:00부터 10:59:59까지 승차한 인원 수 |
| 11~12\_ride | 11:00:00부터 11:59:59까지 승차한 인원 수 |
| 6~7\_takeoff | 6:00:00부터 6:59:59까지 하차한 인원 수 |
| 7~8\_takeoff | 7:00:00부터 7:59:59까지 하차한 인원 수 |
| 8~9\_takeoff | 8:00:00부터 8:59:59까지 하차한 인원 수 |
| 9~10\_takeoff | 9:00:00부터 9:59:59까지 하차한 인원 수 |
| 10~11\_takeoff | 10:00:00부터 10:59:59까지 하차한 인원 수 |
| 11~12\_takeoff | 11:00:00부터 11:59:59까지 하차한 인원 수 |
| 18~20\_ride | 18:00:00부터 19:59:59까지 승차한 인원 수  (train data에만 존재) |

1. bus\_bst.csv

bus\_bts의 경우, 버스카드별로 승하차 정보가 기록이 되어있습니다. 해당 데이터는 탑승 시간대가오전 6시부터 12시 사이인 경우만 있습니다.

|  |  |
| --- | --- |
| 변수명 | 데이터 설명 |
| user\_card\_id | 해당 승객의 버스카드ID |
| bus\_route\_id | 노선ID |
| vhc\_id | 차량ID |
| geton\_date | 해당 승객이 탑승한 날짜 |
| geton\_time | 해당 승객이 탑승한 시간 |
| geton\_station\_code | 승차정류소의 ID |
| geton\_station\_name | 승차정류소의 이름 |
| getoff\_date | 해당 승객이 하차한 날짜 (하차태그 없는 경우, NaN) |
| getoff\_time | 해당 승객이 하차한 시간 (하차태그 없는 경우, NaN) |
| getoff\_station\_code | 하차정류소의 ID (하차태그 없는 경우, NaN) |
| getoff\_station\_name | 하차정류소의 이름 (하차태그 없는 경우, NaN) |
| user\_category | 승객 구분 (하단 표 참고) |
| user\_count | 해당 버스카드로 계산한 인원수  ( ex- 3은 3명 분의 버스비를 해당 카드 하나로 계산한 것) |

bus\_bts.csv의 user\_category (승객 구분 유형)

|  |  |
| --- | --- |
| 승객 구분 번호 | 승객 유형 설명 |
| 01 | 일반 |
| 02 | 어린이 |
| 04 | 청소년 |
| 06 | 경로 |
| 27 | 장애 일반 |
| 28 | 장애 동반 |
| 29 | 유공 일반 |
| 30 | 유공 동반 |

1. submission\_제출양식.csv

submission\_제출양식의 경우, test data의 ID와 목표변수인 18시~20시 승차 인원로만 이루어져 있습니다.

참가자분들께서는 test.csv에서 ID와 나온 예측값을 결합하여, 해당 submission\_제출양식.csv 파일처럼 만든 이후, 제출을 해주시면 됩니다.

|  |  |
| --- | --- |
| 변수명 | 데이터 설명 |
| id | test data에서의 고유한 ID |
| 18~20\_ride | 18:00:00부터 19:59:59까지 승차한 인원 수 |