제주 빅데이터 센터 공공 데이터를 이용한 교통카드 거래기록 데이터의 EDA 분석

2019.9월

제주대학교 SW중심대학사업단 / eINS S&C ㈜ 아인스에스엔씨

교통카드 거래기록 데이터에 대한 EDA

1. 테이블명 : tb_bus_user_usage

① 제공 : 제주 빅데이터 센터

② 제공한 데이터의 수집기간 : 2019년 4월 1일 ~ 2019년8월31일

③ 하루치 데이터의 평균 용량 : 약 30MB

교통카드 거래기록 데이터에 대한 EDA

2. 스키마

칼럼 명	칼럼 타입	설명	비고사항	칼럼 명	칼럼 타입	설명	비고사항
user_id	STRING	사용자 식별자	개인정보 사항으로, 암호화되어 있음.	getoff_datetime	STRING	하차일시	NULL값 다수 존재
base_date	STRING	날짜		getoff_station_id	STRING	하차 정류소 아이디	NULL값 다수 존재
route_id	STRING	버스 노선 아이디		getoff_station_name	DOUBLE	경도	NULL값 다수 존재
route_name	STRING	버스 노선명		getoff_station_longitude	DOUBLE	하차정류소 좌표(경도)	NULL값 다수 존재
route_no	STRING	버스 노선번호		getoff_station_latitude	DOUBLE	하차정류소 좌표(위도)	NULL값 다수 존재
geton_datetime	STRING	승차일시	(예)20190601105559	user_type	STRING	탑승자 유형	일반, 어린이, 청소년, 경로, 장애동반, 장애일반, 유공일반 등
geton_station_id	STRING	승차정류소 아이디		user_count	INT	탑승 인원	(예) 1, 2 , …
geton_station_name	STRING	승차정류소 명		input_date	STRING	입력일	
geton_station_longitude	DOUBLE	승차정류소 좌표(경도)	(예) 126.5271				
geton_station_latitude	DOUBLE	승차정류소 좌표(위도)	(예)33.51656				

교통카드 거래기록 데이터에 대한 EDA

2. 인스턴스 예시

user_id	base_date	route_id	route_name	route_no	geton_datetime	geton_station_id	geton_stataion_name	geton_station_longitude	geton_station_latitude	getoff_datetime	getoff_station_id	getoff_station_name	getoff_station_longitude	getoff_station_latitude	user_type	user_count	input_date
3bfc425c3390890fbf5aa67d359c1f712561fd52c53b71b23b49f77a4527c81a	20190504	24130000	432-1(제주버스터미널~제주버스터미널)	432-1	20190504220545	113	탑동푸른쉼터	126.52711	33.51656			NULL	NULL	NULL	일반	1	20190507
c6e112437023e2fe1d74560a7e055dd2a003aff1151a073c8056cd8cb6428f54	20190504	24130000	432-1(제주버스터미널~제주버스터미널)	432-1	20190504202951	113	탑동푸른쉼터	126.52711	33.51656			NULL	NULL	NULL	일반	1	20190507
63d9fe08935e429d7bbba1921d63dc11247b9bbb55187fae893b4e7a67d2e19a	20190504	17010000	시티투어2(트론리)	시티투어2	20190504143702	6000027	제주웰컴센터	NULL	0			NULL	NULL	NULL	일반	1	20190507
3dbffcf2d82a7fafe1798c2131148234f33b127195cc05b50fa0f122c5284e85	20190504	17010000	시티투어2(트론리)	시티투어2	20190504143714	6000027	제주웰컴센터	NULL	0			NULL	NULL	NULL	일반	1	20190507
cb859f1e5287920538b5d6ff51611d726365a91c0171aa4b499819edba47032b	20190504	17010000	시티투어2(트론리)	시티투어2	20190504105508	6000027	제주웰컴센터	NULL	0			NULL	NULL	NULL	일반	2	20190507
63d9fe08935e429d7bbba1921d63dc11247b9bbb55187fae893b4e7a67d2e19a	20190504	17010000	시티투어2(트론리)	시티투어2	20190504114402	6000027	제주웰컴센터	NULL	0			NULL	NULL	NULL	일반	1	20190507
55c72600744e8195f07a5dcb92e41c8fa8f2628fe5255d9b7bfcaa5e4656ee2c	20190504	22490000	270-2(제주대학교~애월하나로마트)	270-2	20190504114614	3281	남국원(광양방면)	126.54388	33.47957			NULL	NULL	NULL	일반	1	20190507
3dbffcf2d82a7fafe1798c2131148234f33b127195cc05b50fa0f122c5284e85	20190504	17010000	시티투어2(트론리)	시티투어2	20190504114406	6000027	제주웰컴센터	NULL	0			NULL	NULL	NULL	일반	1	20190507
84a62bb8bb5a53ef86fa36dbb255af95b0020d6bde7cbb75dd8f55c5618bdd05	20190504	26180000	635-1(토평마을회관~남주중고등학교)	635-1	20190504095429	1938	주공아파트5단지	126.56115	33.2646			NULL	NULL	NULL	청소년	1	20190507
993c511cc2cfb2cb81b30714971cc02bf8f9cb2b39ea9eff160494eb1dc7069c	20190504	26180000	635-1(토평마을회관~남주중고등학교)	635-1	20190504171219	1938	주공아파트5단지	126.56115	33.2646			NULL	NULL	NULL	청소년	1	20190507
d236f9d8ce103d92c67a8c82f0519e5178f02986ffaf04a4743cc03ec7001028	20190504	26180000	635-1(토평마을회관~남주중고등학교)	635-1	20190504183417	1938	주공아파트5단지	126.56115	33.2646			NULL	NULL	NULL	일반	1	20190507
62d2ae011fd05ac9a6505629998cbf338c06450d7a1e45a72cf96ccdbb2f5826	20190504	22480000	270-1(애월하나로마트~제주대학교)	270-1	20190504201616	1082	애월읍사무소	126.33034	33.46258			NULL	NULL	NULL	일반	1	20190507
ea5460ea38ee4b5ffab44ccb5d591984d8d1c71b0ab3c2827d66aea55879c2b4	20190504	28690000	202-13(제주버스터미널(가상정류소)~서귀포등기소)	202-13	20190504174048	1564	제주시외버스터미널	126.51479	33.49946	20190504175529	95	제주서중학교	126.48135	33.49367	일반	1	20190505
3e97d49fc6cebce5e3308a9b7c0dc3833b0a0a5082222a997d1e7699c798af10	20190504	23620000	365-22(제주한라대학교~제주대학교)	365-22	20190504064215	304	용문사거리	126.51023	33.50866	20190504070501	121	제주대학교병원	126.54739	33.46909	장애 동반	1	20190504
80b4df089f2976e337e3056e07b0ebd104ae79f1449d3eae6e80ca683d654146	20190504	23620000	365-22(제주한라대학교~제주대학교)	365-22	20190504134859	269	삼성초등학교	126.52664	33.5041	20190504140307	121	제주대학교병원	126.54739	33.46909	일반	1	20190504
307ff0af48c5812dfc07dedb3b2f09733a165c3346bd4427bb3f287fd7d7bec9	20190504	23290000	341-2(신사동(기점)~제주대학교)	341-2	20190504132021	268	화북남문	126.56512	33.5189	20190504134908	121	제주대학교병원	126.54739	33.46909	일반	1	20190504

교통카드 거래기록 데이터에 대한 EDA

3. Card_data 오류 사항

Code	오류 유형
1	하차시간 및 하차 정류장 ID 필드 값이 결측값인 오류
2	승차 정류장과 하차 정류장이 일치하는 오류
3	기점(종점)에서 승차(하차) 인원이 존재하는 오류
4	해당 노선이 경유해야 할 정류장과 역순으로 승객들이 승·하차한 것으로 기록된 오류
5	'노선별경유정류장' 데이터간의 오류 (해당 데이터에 노선이나 정류장이 누락됨.)

교통카드 거래기록 데이터에 대한 EDA

[오류 Code1]

하차시간 및 하차 정류장 ID 필드 값이 결측값인 오류

교통카드 거래기록 데이터에 대한 EDA

3-1. 오류 Code1 : 하차시간 및 하차 정류장 ID 필드 값이 결측값인 오류

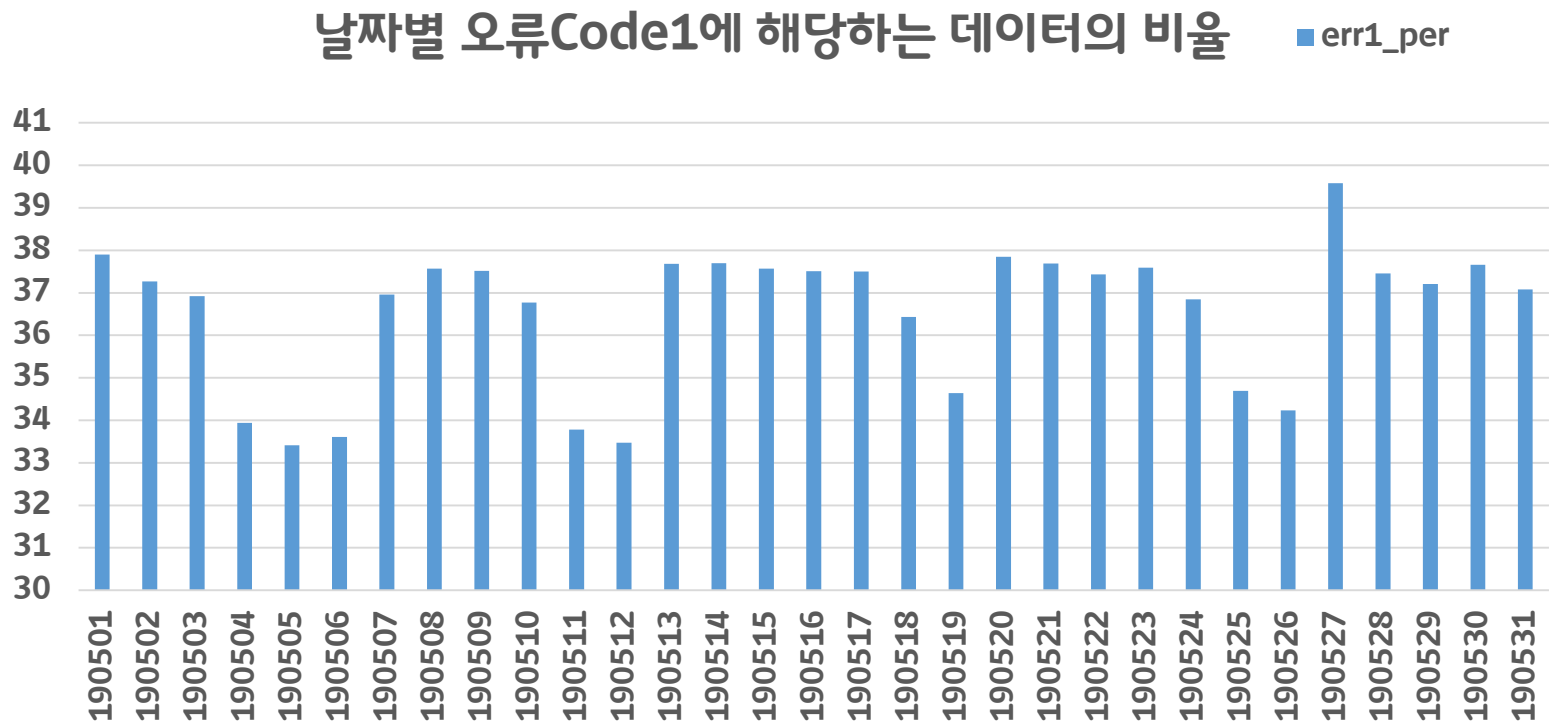
user_id	3bfc425c3390890fbf5aa67d359c1f712561fd52c53b71b23b49f77a4527c81a	geton_datetime	20190504220545	getoff_datetime		user_type	일반
base_date	20190504	geton_station_id	113	getoff_station_id		user_count	1
route_id	24130000	geton_station_name	탑동푸른쉼터	getoff_station_name	NULL	input_date	20190507
route_name	432-1(제주버스터미널~제주버스터미널)	geton_station_longitude	126.52711	getoff_station_longitude	NULL		
route_no	432-1	geton_station_latitude	33.51656	getoff_station_latitude	NULL		

2019/05/04 교통카드 거래 기록 데이터 중 일부

교통카드 거래기록 데이터에 대한 EDA

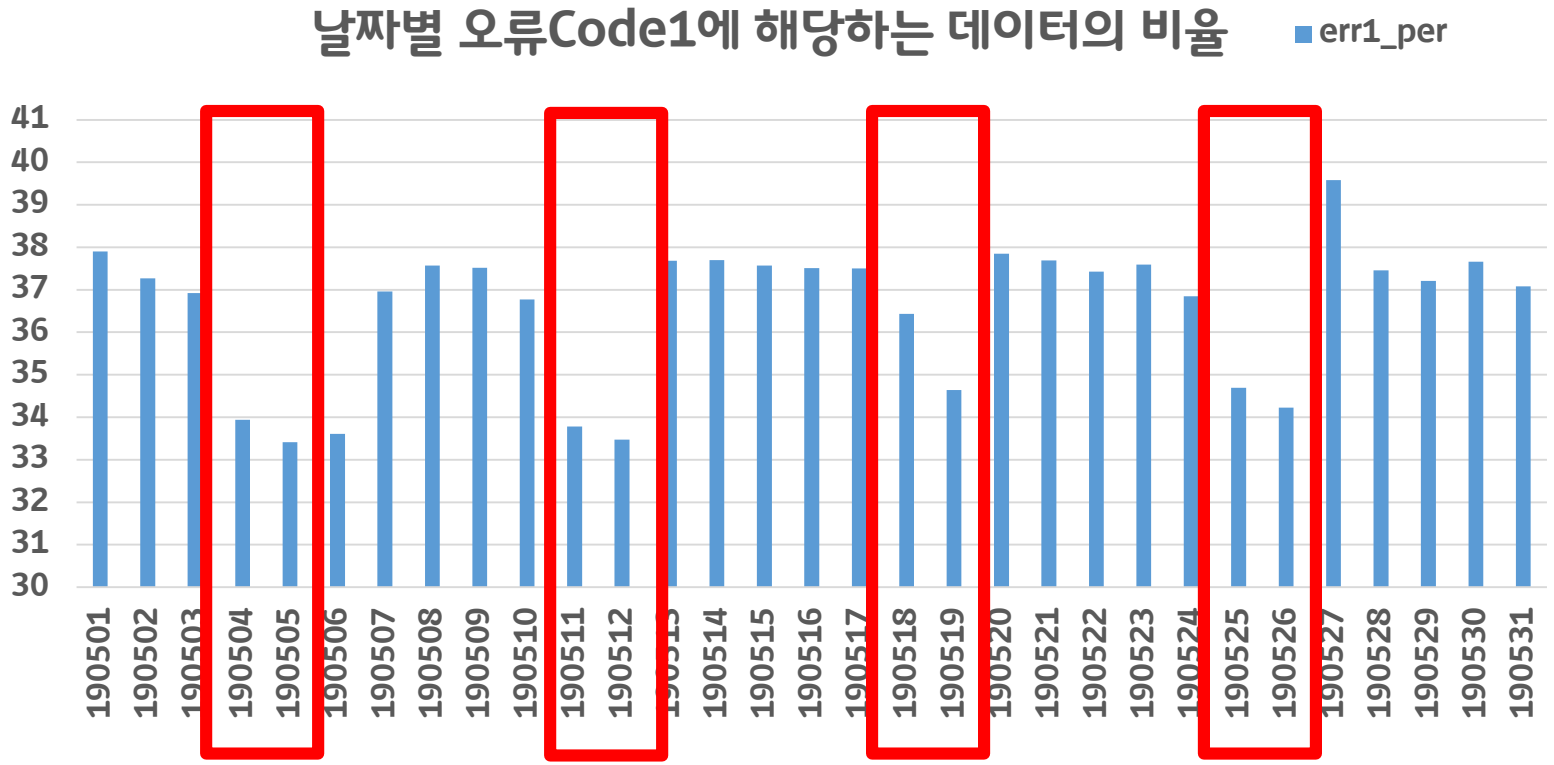
3-1. 오류 Code1 : 하차시간 및 하차 정류장 ID 필드 값이 결측값인 오류

① 5월 교통카드 거래 기록 데이터의 개수 : 4,776,361건
② ERROR CODE1 에 해당하는 데이터 수: 1,756,041건(약36.76%)



교통카드 거래기록 데이터에 대한 EDA

3-1. 오류 Code1 : 하차시간 및 하차 정류장 ID 필드 값이 결측값인 오류



주말에 결측이 적어지는 추세를 볼 수 있음. 이는 평소 학생들이 하차 미태그를 하는 경우가 많은 것으로 예상됨.

[오류 Code2]

승차 정류장과 하차 정류장이 일치하는 오류

교통카드 거래기록 데이터에 대한 EDA

3-2. 오류 Code2: 승차 정류장과 하차 정류장이 일치하는 오류

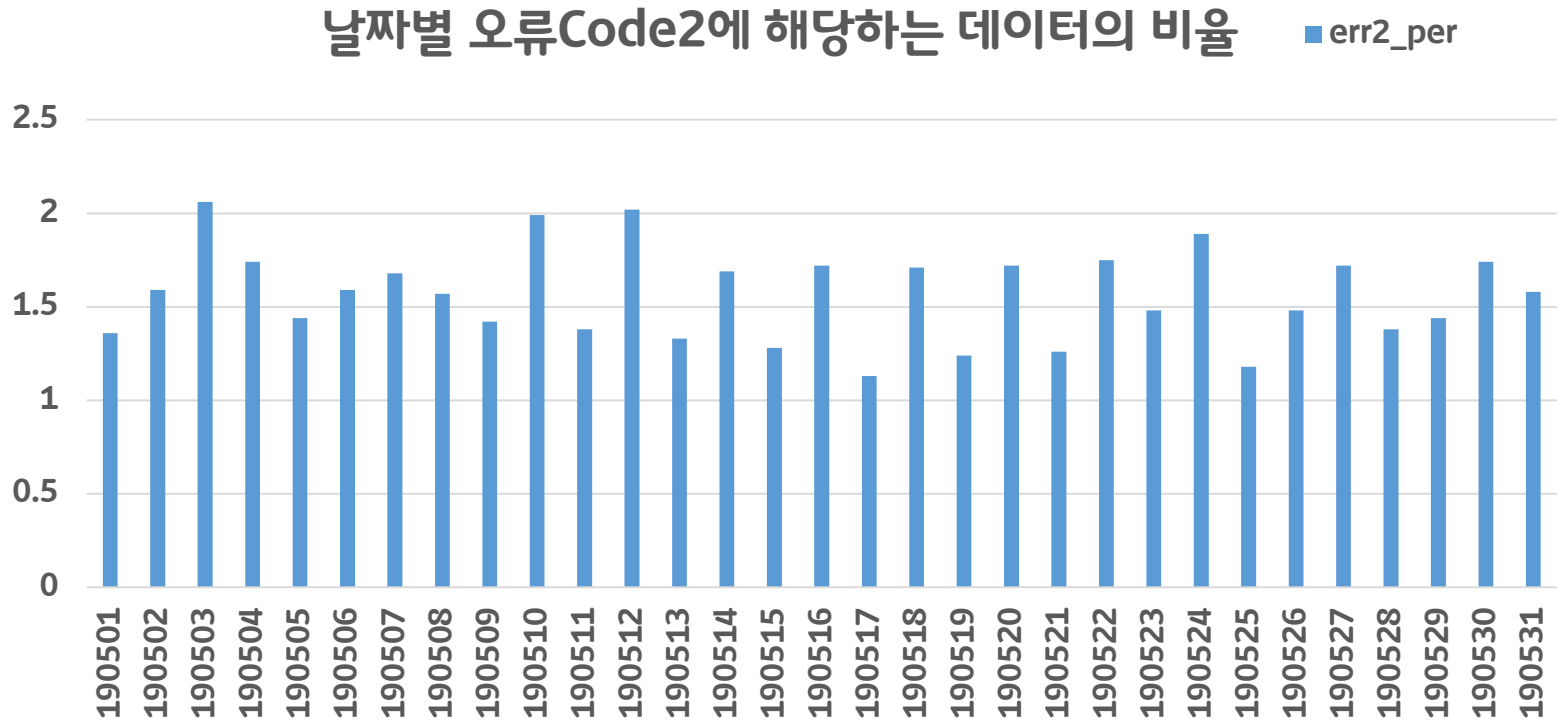
user_id	7207b45dfed39f7dd88791b8093ab3cd86eb2e478378949c8cb2ca7ed00fafe8	geton_datetime	20190504090303	getoff_datetime	20190504090537	user_type	경로
base_date	20190504	geton_station_id	7	getoff_station_id	7	user_count	1
route_id	23140000	geton_station_name	노형주공아파트	getoff_station_name	노형주공아파트	input_date	20190504
route_name	331-1(한라수목원~삼양종점)	geton_station_longitude	126.47418	getoff_station_longitude	126.47418		
route_no	432-1	geton_station_latitude	33.49068	getoff_station_latitude	33.49068		

2019/05/04 교통카드 거래 기록 데이터 중 일부

교통카드 거래기록 데이터에 대한 EDA

3-2. 오류 Code2: 승차 정류장과 하차 정류장이 일치하는 오류

① 5월 교통카드 거래 기록 데이터의 개수 : 4,776,361건
② ERROR CODE1 에 해당하는 데이터 수: 75,001건(약1.57%)



[오류 Code3]

기점(종점)에서 승차(하차) 인원이 존재하는 오류

교통카드 거래기록 데이터에 대한 EDA

3-3. 오류 Code3 : 기점(종점)에서 승차(하차) 인원이 존재하는 오류

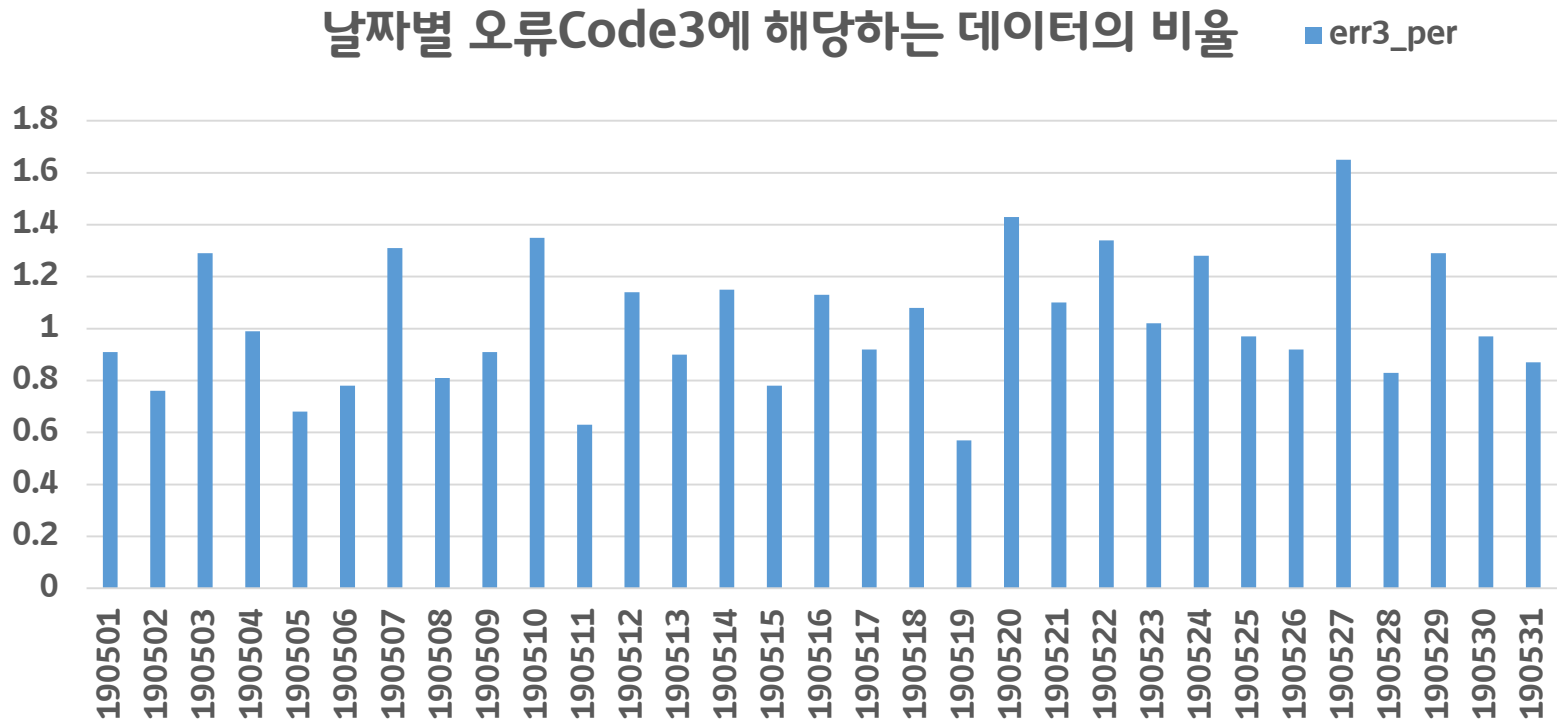
user_id	61319d26de264fd72b56598878c0879b6bab47019cad1355960d916640fdb606	geton_datetime	20190504154422	getoff_datetime	20190504161917	user_type	일반
base_date	20190504	geton_station_id	447	getoff_station_id	149	user_count	1
route_id	23580000	geton_station_name	제주고등학교/중흥S클래스	getoff_station_name	제주버스터미널	input_date	20190504
route_name	360-2(제주대학교~제주고등학교/중흥S클래스)	geton_station_longitude	126.48141	getoff_station_longitude	126.51486		
route_no	360-2	geton_station_latitude	33.4762599999999	getoff_station_latitude	33.49993		

2019/05/04 교통카드 거래 기록 데이터 중 일부

교통카드 거래기록 데이터에 대한 EDA

3-3. 오류 Code3 : 기점(종점)에서 승차(하차) 인원이 존재하는 오류

① 5월 교통카드 거래 기록 데이터의 개수 : 4,776,361건
② ERROR CODE1 에 해당하는 데이터 수: 49,556건(약1.03%)



[오류 Code4]

해당 노선이 경유해야 할 정류장과 역순으로 승객들이 승·하차한 것으로 기록된 오류

교통카드 거래기록 데이터에 대한 EDA

3-4. 오류 Code4 : 해당 노선이 경유해야 할 정류장과 역순으로 승객들이 승·하차한 것으로 기록된 오류

user_id	494dc9f82e3c45e2afb9343a18a55576c420553d04ecfa9dc5c6b3fc05e26fb1	geton_datetime	20190504092123	getoff_datetime	20190504093324	user_type	일반
base_date	20190504	geton_station_id	1613	getoff_station_id	2771	user_count	1
route_id	28150000	geton_station_name	제주유리의성	getoff_station_name	테디밸리골프앤리조트,헬로키티아일랜드	input_date	20190504
route_name	820-1(동광환승정류장1~동광환승정류장(종점))	geton_station_longitude	126.27206	getoff_station_longitude	126.3512		
route_no	820-1	geton_station_latitude	33.31367	getoff_station_latitude	33.2919		

2019/05/04 교통카드 거래 기록 데이터 중 일부

교통카드 거래기록 데이터에 대한 EDA

3-4. 오류 Code4 : 해당 노선이 경유해야 할 정류장과 역순으로 승객들이 승·하차한 것으로 기록된 오류

route_name	820-1(동광환승정류장1~동광환승정류장(종점))
geton_station_name	제주유리의성
getoff_station_name	테디밸리골프앤리조트,헬로키티아일랜드

2019/05/04
교통카드 거래 기록 데이터 중 일부

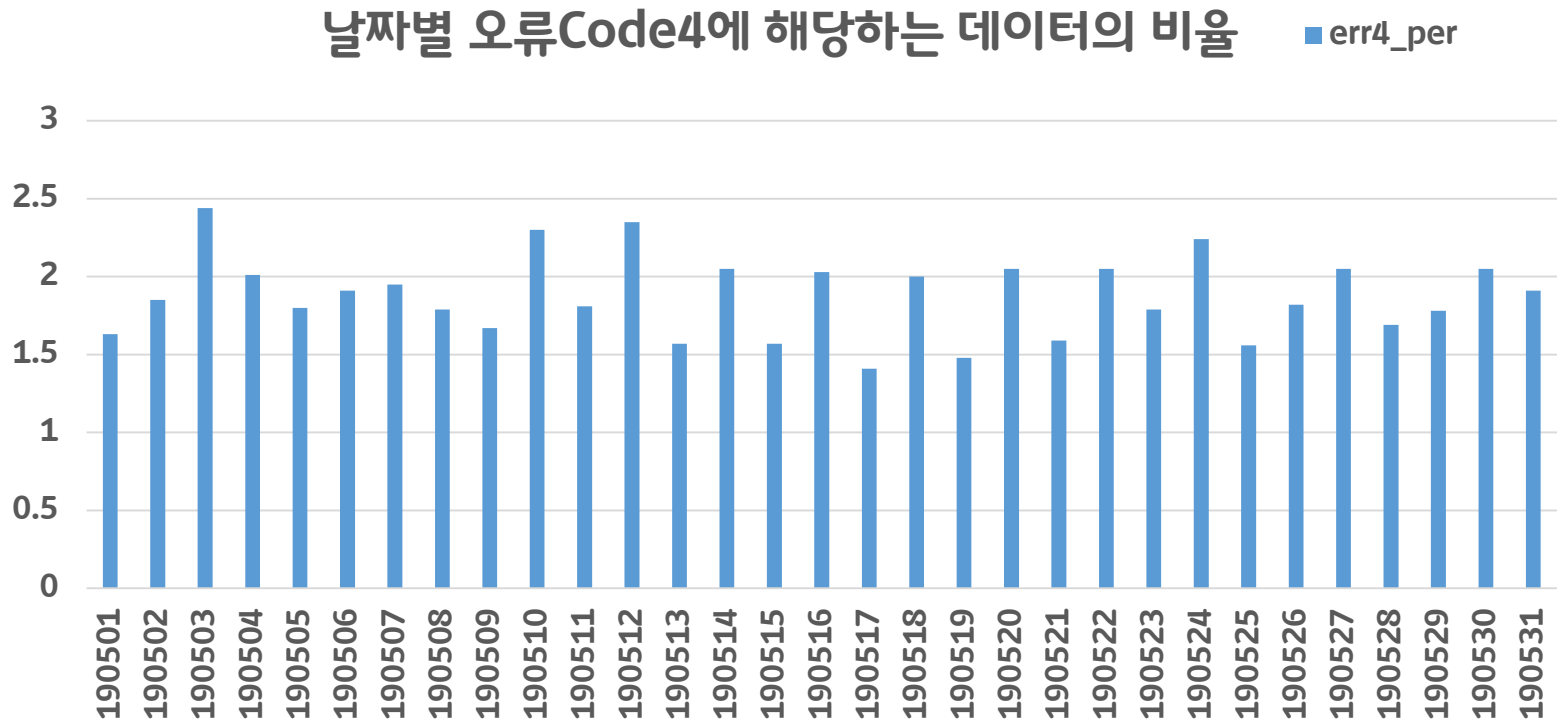
1	동광환승정류장1
2	테디밸리골프앤리조트,헬로키티아일랜드
3	세계자동차제주박물관
…	중략
20	제주유리의성
21	제주오설록 티뮤지엄

820-1 노선의 경유정류장 정보

교통카드 거래기록 데이터에 대한 EDA

3-4. 오류 Code4 : 해당 노선이 경유해야 할 정류장과 역순으로 승객들이 승·하차한 것으로 기록된 오류

① 5월 교통카드 거래 기록 데이터의 개수 : 4,776,361건
② ERROR CODE1 에 해당하는 데이터 수: 89,789건(약1.87%)



[오류 Code5]

'노선별경유정류장' 데이터의 누락 오류
(해당 데이터에 노선이나 정류장이 누락됨.)

교통카드 거래기록 데이터에 대한 EDA

3-5. 오류 Code5 : '노선별경유정류장' 데이터의 누락 오류(해당 데이터에 노선이나 정류장이 누락됨.)

user_id	07824a814c01cc8dd6d3daaf4cd715aaccf5a9b396a19c292780aece99cabda6	geton_datetime	20190504144104	getoff_datetime	20190504154320	user_type	일반
base_date	20190504	geton_station_id	1355	getoff_station_id	320	user_count	1
route_id	23210000	geton_station_name	제주국제공항(신제주방면)	getoff_station_name	제주도청신제주로터리	input_date	20190504
route_name	332-4(삼양3동포구~한라수목원)	geton_station_longitude	126.49275	getoff_station_longitude	126.49678		
route_no	332-4	geton_station_latitude	33.5061	getoff_station_latitude	33.49143		

2019/05/04 교통카드 거래 기록 데이터 중 일부

교통카드 거래기록 데이터에 대한 EDA

3-5. 오류 Code5 : '노선별경유정류장' 데이터의 누락 오류(해당 데이터에 노선이나 정류장이 누락됨.)

route_name	332-4(삼양3동포구~한라수목원)
geton_station_name	제주국제공항(신제주방면)
geton_station_id	1355
getoff_station_name	제주도청신제주로터리
getoff_station_id	320

2019/05/04
교통카드 거래 기록 데이터 중 일부

순서	정류장명	정류장번호
33	월성마을	392
34	제주국제공항(신제주방면)	1355
35	다호마을	620
…	중략	…
37	제주도청신제주로터리	321
38	수협제주도지회	345

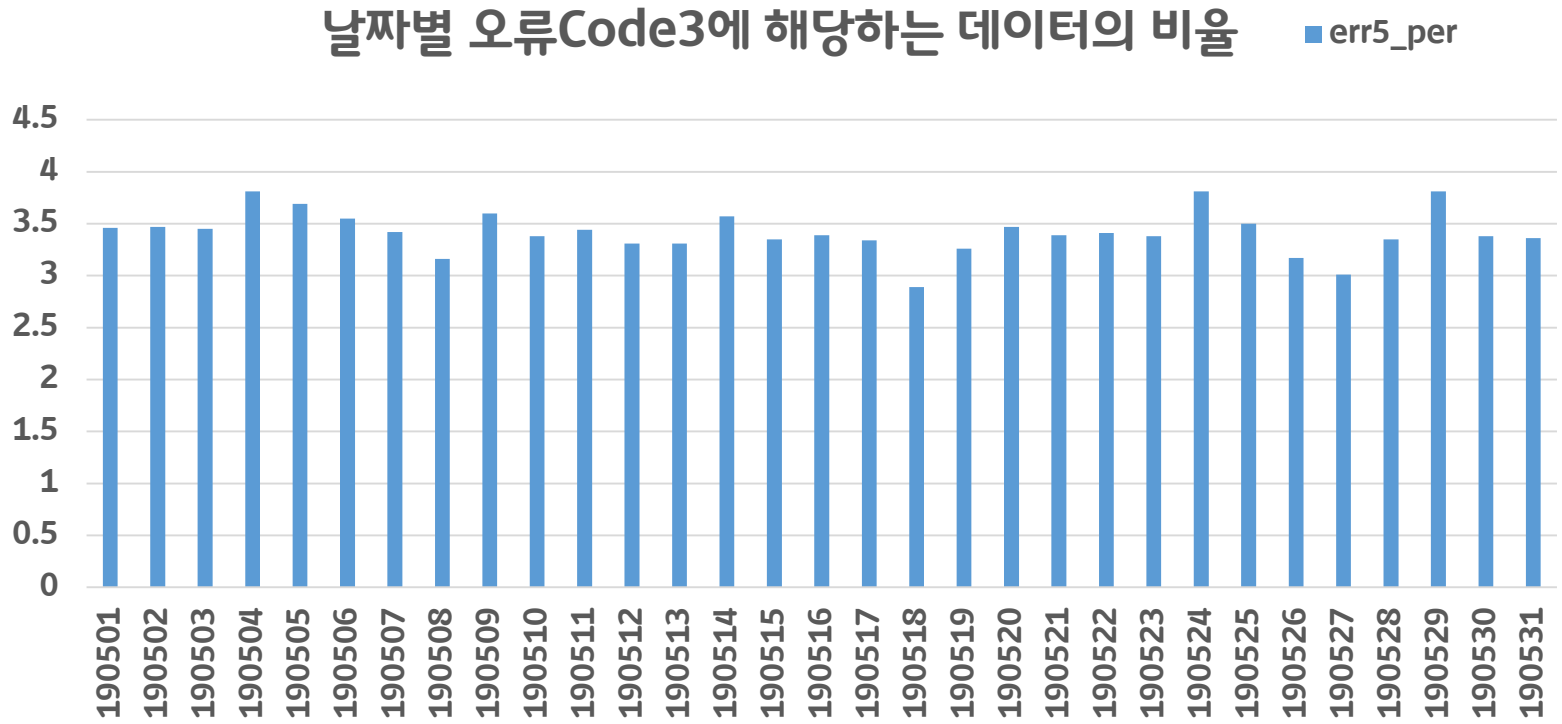
332-4 노선의 경유정류장 정보

=> "제주도청신제주로터리 " 라는 정류장명은 동일하지만, 정류장 번호가 다르기 때문에 다른 정류장임.

교통카드 거래기록 데이터에 대한 EDA

3-5. 오류 Code5 : '노선별경유정류장' 데이터의 누락 오류(해당 데이터에 노선이나 정류장이 누락됨.)

① 5월 교통카드 거래 기록 데이터의 개수 : 4,776,361건
② ERROR CODE1 에 해당하는 데이터 수: 163,558건(약3.42%)



교통카드 거래기록 데이터에 대한 EDA

4. Card_data 오류 사항 별 정제/보정 방안

Code	오류 유형
1	하차시간 및 하차 정류장 ID 필드 값이 결측값인 오류
2	승차 정류장과 하차 정류장이 일치하는 오류
3	기점(종점)에서 승차(하차) 인원이 존재하는 오류
4	해당 노선이 경유해야 할 정류장과 역순으로 승객들이 승·하차한 것으로 기록된 오류
5	'노선별경유정류장' 데이터간의 오류 (해당 데이터에 노선이나 정류장이 누락됨.)

교통카드 거래기록 데이터에 대한 EDA

4-1. 오류 Code1 : 하차시간 및 하차 정류장 ID 필드 값이 결측값인 오류

① 단순제거
② 선형보정
③ 가중치 보정
④ 머신러닝을 이용한 하차 정류장 예측 보정 방법
⑤ 통행사슬모형을 사용한 하차 정류장 예측 보정 방법

=> 현실적으로 " ① 단순제거 " 가 가장 타당하다보임.

교통카드 거래기록 데이터에 대한 EDA

4-2. 오류 Code2: 승차 정류장과 하차 정류장이 일치하는 오류

=> 승차/하차 정보 중 어떠한 데이터가 이상치 인지 알 수 없으므로, 단순 제거함.

4-3. 오류 Code3 : 기점(종점)에서 승차(하차) 인원이 존재하는 오류

=> 정확한 보정방법을 꼬할 수 없을 뿐만 아니라, 기계 오작동 또는 기사님의 작동 미숙으로 잠정 결론 내렸기 때문에, 단순 제거함.

4-4~5. 오류 Code4,5

=> 현재로써, '노선별경유정류장' 데이터가 업데이트가 되지 않아 생긴 문제로 판단되어, 단순 제거함. 추후 더 좋은 방안이 마련되는 대로 보정 예정임.