

## 3-005-280 제주도 및 도서지역 데이터

인공지능 학습용 데이터 구축·활용 가이드라인	사업 총괄	(주)데이터웨이
	데이터 설계	(주)데이터웨이
	데이터 획득(수집)	(주)케이스태리서치
	데이터 정제	(주)올포랜드
	데이터 가공	(주)지디에스컨설팅그룹
	데이터 검사	(주)데이터웨이
	클라우드 소싱	에이드리븐(주)
	저작도구 개발	(주)지디에스컨설팅그룹
	AI모델 개발	고려대학교산학협력단
데이터 구축·활용 가이드라인 작성	(주)데이터웨이	김정남
데이터 구축·활용 가이드라인 버전	ver 0.1 ('23. 01. 30)	

[illegible]

## 목 차

<b>제1장 데이터 명세</b>	<b>1</b>
1. 데이터 정보 요약	1
2. 데이터 설계	2
3. 데이터 포맷	3
4. 데이터 구성	10
5. 데이터 통계	13
6. 원시데이터 특성	14
7. 기타 정보	16
<b>제2장 데이터 구축</b>	<b>18</b>
1. 데이터 구축 개요	18
2. 임무 정의	20
3. 획득(수집)	21
4. 정제	29
5. 가공	34
6. 검사	38
7. 학습 모델	47
<b>제3장 데이터 활용</b>	<b>53</b>
1. 데이터 활용	53
2. 응용 서비스	54
3. 응용 서비스 개발	55
4. 기술 지원	56

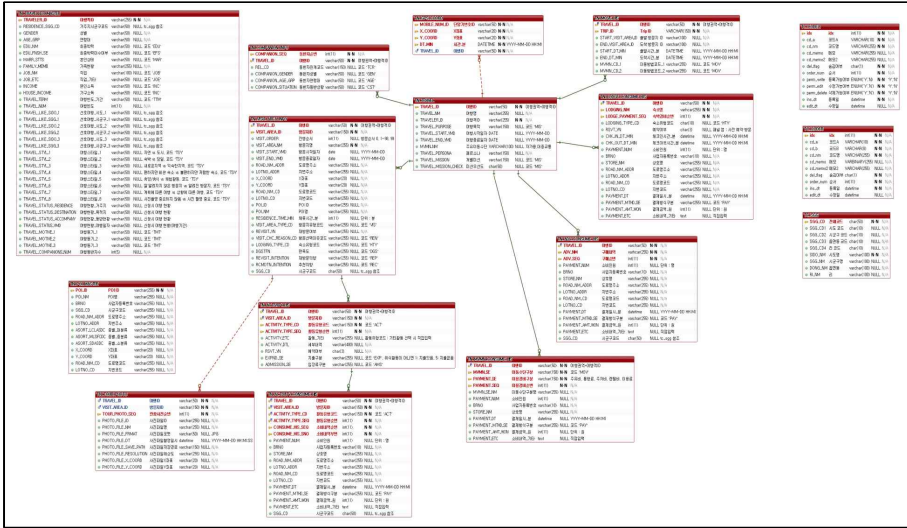
## 제1장 데이터 명세 정보

### 1. 데이터 정보 요약

데이터 명	[3-005-277] 국내 여행로그 데이터(수도권) [3-005-278] 국내 여행로그 데이터(동부권) [3-005-279] 국내 여행로그 데이터(서부권) [3-005-280] 국내 여행로그 데이터(제주도 및 도서지역)						
활용 분야	관광						
데이터 요약	여행객(16,000명)들을 대상으로 여행로그 데이터 구축 - 국내를 수도권/동부권/서부권/제주및도서권 등 4개의 권역으로 나누어 각각 4,000명 씩 총 16,000명의 여행객들을 대상으로 활동내역, 방문지, 소비내역 등의 데이터를 구축						
데이터 출처	여행자 전용 앱(여행로그 앱)을 통하여 데이터 수집						
데이터 통계	데이터 구축 규모	여행자 정보 : 여행자 패널 데이터 (CSV)				16,000 세트	
		동선정보 : GPS 데이터 (CSV)				16,000 세트	
		활동정보 : 여행기록 데이터 (CSV)				16,000 세트	
		여행지 사진 데이터 (JPG)				726,522 장	
		소비내역 데이터 (CSV)				16,000 세트	
		POI 데이터 (CSV)				1식	
	데이터 분포 (충분성, 균등성, 편향성 여부 확인)	성별	남	41%	39%	38%	37%
			여	59%	61%	62%	63%
		연령별	20대	35%	33%	35%	35%
			30대	36%	33%	34%	34%
40대			16%	16%	15%	16%	
50대 ↑			14%	17%	16%	15%	
여행기간별	당일	55%	45%	47%	13%		
	1박 2일	31%	38%	39%	23%		
	2박 3일	14%	17%	14%	63%		
데이터 이력	배포 버전	v0.1					
	개정 이력	신규					
	작성자/배포자	김정남/(주)데이터웨이					

## 2. 데이터 설계

### DB 스키마 정의(ERD)



## 3. 데이터 포맷

한글데이터블명		여행객 Master	영문데이터블명	TN_TRAVELLER_MASTER		
NO	컬럼ID	컬럼명	타입(길이)	NULL	KEY	비고
1	TRAVELER_ID	여행객ID	varchar(255)	N	PK	
2	RESIDENCE_SGG_CODE	거주지시군구코드	varchar(50)	Y		tc_sgg 참조
3	GENDER	성별	varchar(50)	Y		
4	AGE_GRP	연령대	varchar(50)	Y		
5	EDU_NM	최종학력	varchar(50)	Y		코드 'EDU'
6	EDU_FNSH_SE	최종학력이수여부	varchar(50)	Y		코드 'EFS'
7	MARR_STS	혼인상태	varchar(50)	Y		코드 'MAR'
8	FAMILY_MEMB	가족현황	varchar(255)	Y		
9	JOB_NM	직업	varchar(100)	Y		코드 'JOB'
10	JOB_ETC	직업_기타	varchar(50)	Y		코드 'JOE'
11	INCOME	본인소득	varchar(50)	Y		코드 'INC'
12	HOUSE_INCOME	가구소득	varchar(50)	Y		코드 'INC'
13	TRAVEL_TERM	여행빈도_기간	varchar(50)	Y		코드 'TTM'
14	TRAVEL_NUM	여행빈도	int(11)	Y		
15	TRAVEL_LIKE_SIDO_1	선호여행_시도_1	varchar(50)	Y		tc_sgg 참조
16	TRAVEL_LIKE_SGG_1	선호여행_시군구_1	varchar(50)	Y		tc_sgg 참조
17	TRAVEL_LIKE_SIDO_2	선호여행_시도_2	varchar(50)	Y		tc_sgg 참조
18	TRAVEL_LIKE_SGG_2	선호여행_시군구_2	varchar(50)	Y		tc_sgg 참조
19	TRAVEL_LIKE_SIDO_3	선호여행_시도_3	varchar(50)	Y		tc_sgg 참조
20	TRAVEL_LIKE_SGG_3	선호여행_시군구_3	varchar(50)	Y		tc_sgg 참조
21	TRAVEL_STYL_1	여행스타일_1	varchar(50)	Y		코드 'TSY'
22	TRAVEL_STYL_2	여행스타일_2	varchar(50)	Y		코드 'TSY'
23	TRAVEL_STYL_3	여행스타일_3	varchar(50)	Y		코드 'TSY'
24	TRAVEL_STYL_4	여행스타일_4	varchar(50)	Y		코드 'TSY'
25	TRAVEL_STYL_5	여행스타일_5	varchar(50)	Y		코드 'TSY'
26	TRAVEL_STYL_6	여행스타일_6	varchar(50)	Y		코드 'TSY'
27	TRAVEL_STYL_7	여행스타일_7	varchar(50)	Y		코드 'TSY'
28	TRAVEL_STYL_8	여행스타일_8	varchar(50)	Y		코드 'TSY'
29	TRAVEL_STATUS_RESIDENCE	여행현황_거주지	varchar(50)	Y		
30	TRAVEL_STATUS_DESTINATION	여행현황_목적지	varchar(50)	Y		
31	TRAVEL_STATUS_ACCOMPANY	여행현황_동반현황	varchar(50)	Y		
32	TRAVEL_STATUS_YMD	여행현황_여행일자	varchar(50)	Y		
33	TRAVEL_MOTIVE_1	여행동기_1	varchar(50)	Y		코드 'TMT'
34	TRAVEL_MOTIVE_2	여행동기_2	varchar(50)	Y		코드 'TMT'
35	TRAVEL_MOTIVE_3	여행동기_3	varchar(50)	Y		코드 'TMT'
36	TRAVEL_COMPANIONS_NUM	여행동반자수	int(5)	Y		

한글테이블명		POI Master		영문테이블명		TN_POI_MASTER	
NO	컬럼ID	컬럼명	타입	NULL	KEY	비고	
1	POI_ID	POI ID	varchar(255)	N	PK		
2	POI_NM	POI명	varchar(255)	Y			
3	BRNO	사업자등록번호	varchar(255)	Y			
4	SGG_CD	시군구코드	varchar(255)	Y			
5	ROAD_NM_ADDR	도로명주소	varchar(255)	Y			
6	LOTNO_ADDR	지번주소	varchar(255)	Y			
7	ASORT_LCLASDC	종별_대분류	varchar(255)	Y			
8	ASORT_MLSFCD	종별_중분류	varchar(255)	Y			
9	ASORT_SDASDC	종별_소분류	varchar(255)	Y			
10	X_COORD	X좌표	varchar(20)	Y			
11	Y_COORD	Y좌표	varchar(20)	Y			
12	ROAD_NM_CD	도로명코드	varchar(255)	Y			
13	LOTNO_CD	지번코드	varchar(255)	Y			

한글테이블명		여행		영문테이블명		TN_TRAVEL	
NO	컬럼ID	컬럼명	타입	NULL	KEY	비고	
1	TRAVEL_ID	여행ID	varchar(50)	N	PK		
2	TRAVEL_NM	여행명	varchar(255)	Y			
3	TRAVELER_ID	여행객ID	varchar(255)	N			
4	TRAVEL_PURPOSE	여행목적	varchar(150)	Y		코드 'MIS'	
5	TRAVEL_START_YMD	여행시작일자	date	Y		YYYY-MM-DD	
6	TRAVEL_END_YMD	여행종료일자	date	Y		YYYY-MM-DD	
7	MVMN_NM	주요이동수단	varchar(100)	Y			
8	TRAVEL_PERSONA	페르소나	varchar(150)	Y			
9	TRAVEL_MISSION	개별미션	varchar(150)	Y		코드 'MIS'	
10	TRAVEL_MISSION_CHECK	미션우선도	char(50)			코드 'MIS'	

한글테이블명		동반자정보		영문테이블명		TN_COMPANION_INFO	
NO	컬럼ID	컬럼명	타입	NULL	KEY	비고	
1	COMPANION_SEQ	동반자순번	int(11)	N	PK		
2	TRAVEL_ID	여행ID	varchar(50)	N	PK,FK		
3	REL_CD	동반자관계코드	varchar(150)	Y		코드 'TCR'	
4	COMPANION_GENDER	동반자성별	varchar(50)	Y		코드 'GEN'	
5	COMPANION_AGE_GRP	동반자연령대	varchar(50)	Y		코드 'AGE'	
6	COMPANION_SITUATION	동반자동반상황	varchar(50)	Y		코드 'CST'	

한글테이블명		이동내역		영문테이블명		TN_MOVE_HIS	
NO	컬럼ID	컬럼명	타입	NULL	KEY	비고	
1	TRIP_ID	Trip ID	varchar(150)	N	PK		
2	TRAVEL ID	여행ID	varchar(50)	N	PK,FK		

한글테이블명		이동내역		영문테이블명		TN_MOVE_HIS	
NO	컬럼ID	컬럼명	타입	NULL	KEY	비고	
3	START_VISIT_AREA_ID	출발 방문지 ID	varchar(100)	Y			
4	END_VISIT_AREA_ID	도착 방문지 ID	varchar(100)	Y			
5	START_DT_MIN	출발시간_분	datetime	Y		YYYY-MM-DD HH:MI	
6	END_DT_MIN	도착시간_분	datetime	Y		YYYY-MM-DD HH:MI	
7	MVMN_CD_1	이동방법코드_1	varchar(255)	Y		코드 'MOV'	
8	MVMN_CD_2	이동방법코드_2	varchar(255)	Y		코드 'MOV'	

한글테이블명		GPS좌표		영문테이블명		TN_GPS_COORD	
NO	컬럼ID	컬럼명	타입	NULL	KEY	비고	
1	MOBILE_NUM_ID	단말기번호ID	varchar(50)	N	PK		
2	X_COORD	X좌표	varchar(20)	N	PK		
3	Y_COORD	Y좌표	varchar(20)	N	PK		
4	DT_MIN	시간_분	datetime	N	PK	YYYY-MM-DD HH:MI	
5	TRAVEL_ID	여행ID	varchar(50)	N	PK,FK		

한글테이블명		이동수단소비내역		영문테이블명		TN_MVMN_CONSUME_HIS	
NO	컬럼ID	컬럼명	타입	NULL	KEY	비고	
1	TRAVEL_ID	여행ID	varchar(50)	N	PK,FK		
2	MVMN_SE	이동수단구분	varchar(150)	N	PK	코드 'MOV'	
3	PAYMENT_SE	이용경비구분	varchar(150)	N	PK		
4	PAYMENT_SEQ	이용경비순번	int(11)	N	PK		
5	MVMN_SE_NM	이동수단구분명	varchar(255)	Y			
6	PAYMENT_NUM	포함인원	int(11)	Y			
7	BRNO	사업자등록번호	varchar(10)	Y			
8	STORE_NM	상호명	varchar(255)	Y			
9	PAYMENT_DT	결제일시_분	datetime	Y		YYYY-MM-DD HH:MI	
10	PAYMENT_MTHD_SE	결제방식구분	varchar(255)	Y		코드 'PAY'	
11	PAYMENT_AMT_WON	결제금액_원	int(11)	Y			
12	PAYMENT_ETC	소비내역_기타	text	Y			

한글테이블명		숙박소비내역		영문테이블명		TN_LODGE_CONSUME_HIS	
NO	컬럼ID	컬럼명	타입	NULL	KEY	비고	
1	TRAVEL_ID	여행ID	varchar(50)	N	PK,FK		
2	LODGING_NM	숙소명	varchar(255)	N	PK		
3	LODGE_PAYMENT_SEQ	숙박경비순번	int(11)	N	PK		
4	LODGING_TYPE_CD	숙소유형코드	char(10)	Y		코드 'HTY'	
5	RSVT_YN	예약여부	char(3)	Y			

한글테이블명		숙박소비내역		영문테이블명		TN_LODGE_CONSUME_HIS
NO	컬럼ID	컬럼명	타입	NULL	KEY	비고
6	CHK_IN_DT_MIN	체크인시간_분	datetime	Y		YYYY-MM-DD HH:MI
7	CHK_OUT_DT_MIN	체크아웃시간_분	datetime	Y		YYYY-MM-DD HH:MI
8	PAYMENT_NUM	소비인원	int(11)	Y		단위 : 명
9	BRNO	사업자등록번호	varchar(10)	Y		
10	STORE_NM	상호명	varchar(255)	Y		
11	ROAD_NM_ADDR	도로명주소	varchar(255)	Y		
12	LOTNO_ADDR	지번주소	varchar(255)	Y		
13	ROAD_NM_CD	도로명코드	varchar(255)	Y		
14	LOTNO_CD	지번코드	varchar(255)	Y		
15	PAYMENT_DT	결제일시_분	datetime	Y		YYYY-MM-DD HH:MI
16	PAYMENT_MTHD_SE	결제방식구분	varchar(255)	Y		코드 'PAY'
17	PAYMENT_AMT_WON	결제금액_원	int(11)	Y		
18	PAYMENT_ETC	소비내역_기타	text	Y		

한글테이블명		사전소비내역		영문테이블명		TN_ADV_CONSUME_HIS
NO	컬럼ID	컬럼명	타입	NULL	KEY	비고
1	TRAVEL_ID	여행ID	varchar(50)	N	PK,FK	
2	ADV_NM	구매내역	varchar(255)	N	PK	
3	ADV_SEQ	구매순번	int(11)	N	PK	
4	PAYMENT_NUM	소비인원	int(11)	Y		단위 : 명
5	BRNO	사업자등록번호	varchar(10)	Y		
6	STORE_NM	상호명	varchar(255)	Y		
7	ROAD_NM_ADDR	도로명주소	varchar(255)	Y		
8	LOTNO_ADDR	지번주소	varchar(255)	Y		
9	ROAD_NM_CD	도로명코드	varchar(255)	Y		
10	LOTNO_CD	지번코드	varchar(255)	Y		
11	PAYMENT_DT	결제일시_분	datetime	Y		YYYY-MM-DD HH:MI
12	PAYMENT_MTHD_SE	결제방식구분	varchar(255)	Y		코드 'PAY'
13	PAYMENT_AMT_WON	결제금액_원	int(11)	Y		
14	PAYMENT_ETC	소비내역_기타	text	Y		
15	SGG_CD	시군구코드	char(50)	Y		tc_sgg 참조

한글테이블명		방문지정보		영문테이블명		TN_VISIT_AREA_INFO
NO	컬럼ID	컬럼명	타입	NULL	KEY	비고
1	TRAVEL_ID	여행ID	varchar(50)	N	PK,FK	
2	VISIT_AREA_ID	방문지 ID	varchar(150)	N	PK	
3	VISIT_ORDER	진행순서	int(11)	N		
4	VISIT_AREA_NM	방문지명	varchar(255)	N		
5	VISIT_START_YMD	방문시작일자	date	Y		YYYY-MM-D

한글테이블명		방문지정보		영문테이블명		TN_VISIT_AREA_INFO
NO	컬럼ID	컬럼명	타입	NULL	KEY	비고
						D
6	VISIT_END_YMD	방문종료일자	date	Y		YYYY-MM-DD
7	ROAD_NM_ADDR	도로명주소	varchar(255)	Y		
8	LOTNO_ADDR	지번주소	varchar(255)	Y		
9	X_COORD	X좌표	varchar(20)	Y		
10	Y_COORD	Y좌표	varchar(20)	Y		
11	ROAD_NM_CD	도로명코드	varchar(255)	Y		
12	LOTNO_CD	지번코드	varchar(255)	Y		
13	POI_ID	POI ID	varchar(255)	Y		
14	POI_NM	POI명	varchar(255)	Y		
15	RESIDENCE_TIME_MIN	체류시간_분	int(11)	Y		단위 : 분
16	VISIT_AREA_TYPE_CD	방문지유형코드	varchar(255)	Y		코드 'VIS'
17	REVISIT_YN	재방문여부	varchar(255)	Y		
18	VISIT_CHC_REASON_CD	방문선택이유코드	varchar(255)	Y		코드 'REN'
19	LODGING_TYPE_CD	숙소유형코드	varchar(255)	Y		코드 'HTY'
20	DGSTFN	만족도	varchar(255)	Y		코드 'DGS'
21	REVISIT_INTENTION	재방문의향	varchar(255)	Y		코드 'REP'
22	RCMDTN_INTENTION	추천의향	varchar(255)	Y		코드 'REC'
23	SGG_CD	시군구코드	char(50)	Y		tc_sgg 참조

한글테이블명		관광사진		영문테이블명		TN_TOUR_PHOTO
NO	컬럼ID	컬럼명	타입	NULL	KEY	비고
1	TRAVEL_ID	여행ID	varchar(50)	N	FK	
2	VISIT_AREA_ID	방문지 ID	varchar(150)	N	FK	
3	TOUR_PHOTO_SEQ	관광사진순번	int(11)	N	PK	
4	PHOTO_FILE_ID	사진파일ID	varchar(255)	Y		
5	PHOTO_FILE_NM	사진파일명	varchar(255)	Y		
6	PHOTO_FILE_FRMAT	사진파일포맷	varchar(50)	Y		JPG
7	PHOTO_FILE_DT	사진파일촬영일시	datetime	Y		YYYY-MM-DD HH:MI:SS
8	PHOTO_FILE_SAVE_PATH	사진파일저장경로	varchar(150)	Y		
9	PHOTO_FILE_RESOLUTION	사진파일해상도	varchar(255)	Y		
10	PHOTO_FILE_X_COORD	사진파일X좌표	varchar(20)	Y		
11	PHOTO_FILE_Y_COORD	사진파일Y좌표	varchar(20)	Y		

한글테이블명		활동내역		영문테이블명		TN_ACTIVITY_HIS
NO	컬럼ID	컬럼명	타입	NULL	KEY	비고
1	TRAVEL_ID	여행ID	varchar(50)	N	PK,FK	
2	VISIT_AREA_ID	방문지 ID	varchar(150)	N	PK,FK	
3	ACTIVITY_TYPE_CD	활동유형코드	varchar(150)	N	PK	코드 'ACT'
4	ACTIVITY_TYPE_SEQ	활동유형순번	int(11)	N	PK	

한글테이블명		활동내역		영문테이블명		TN_ACTIVITY_HIS	
NO	컬럼ID	컬럼명	타입	NULL	KEY	비고	
5	ACTIVITY_ETC	활동_기타	varchar(255)	Y		직접입력	
6	ACTIVITY_DTL	세부내역	varchar(400)	Y			
7	RSVT_YN	예약여부	char(3)	Y			
8	EXPND_SE	지출구분	varchar(255)	Y		코드 'EXP'	
9	ADMISSION_SE	입장료구분	varchar(255)	Y		코드 'AMS'	

한글테이블명		활동소비내역		영문테이블명		TN_ACTIVITY_CONSUME_HIS
NO	컬럼ID	컬럼명	타입	NULL	KEY	비고
1	TRAVEL_ID	여행ID	varchar(50)	N	PK,FK	
2	VISIT_AREA_ID	방문지 ID	varchar(150)	N	PK,FK	
3	ACTIVITY_TYPE_CD	활동유형코드	varchar(150)	N	PK,FK	코드 'ACT'
4	ACTIVITY_TYPE_SEQ	활동유형순번	int(11)	N	PK,FK	
5	CONSUME_HIS_SEQ	소비내역순번	int(11)	N	PK	
6	CONSUME_HIS_SNO	소비내역부번	int(11)	N	PK	
7	PAYMENT_NUM	소비인원	int(11)	Y		단위 : 명
8	BRNO	사업자등록번호	varchar(10)	Y		
9	STORE_NM	상호명	varchar(255)	Y		
10	ROAD_NM_ADDR	도로명주소	varchar(255)	Y		
11	LOTNO_ADDR	지번주소	varchar(255)	Y		
12	ROAD_NM_CD	도로명코드	varchar(255)	Y		
13	LOTNO_CD	지번코드	varchar(255)	Y		
14	PAYMENT_DT	결제일시_분	datetime	Y		YYYY-MM-DD HH:MI
15	PAYMENT_MTHD_SE	결제방식구분	varchar(255)	Y		코드 'PAY'
16	PAYMENT_AMT_WON	결제금액_원	int(11)	Y		
17	PAYMENT_ETC	소비내역_기타	text	Y		
18	SGG_CD	시군구코드	char(50)	Y		tc_sgg 참조

한글테이블명		시군구		영문테이블명		TC_SGG	
NO	컬럼ID	컬럼명	타입	NULL	KEY	비고	
1	SGG_CD	전체코드	char(50)	N	PK		
2	SGG_CD1	시도코드	char(10)	Y			
3	SGG_CD2	시군구코드	char(10)	Y			
2	SGG_CD3	읍면동코드	char(10)	Y			
3	SGG_CD4	리코드	char(10)	Y			
2	SIDO_NM	시도명	varchar(100)	N			
3	SGG_NM	시군구명	varchar(100)	Y			
2	DONG_NM	읍면동	varchar(100)	Y			
3	RI_NM	리	varchar(100)	Y			

한글테이블명		코드A		영문테이블명		TC_CODEA	
NO	컬럼ID	컬럼명	타입	NULL	KEY	비고	
1	idx	idx	int(11)	N	PK		
2	cd_a	코드A	varchar(10)	N			
3	cd_nm	코드A명	varchar(255)	N			
4	cd_memo	메모	varchar(255)	Y			
5	cd_memo2	메모2	varchar(255)	Y			
6	del_flag	숨김여부	char(1)	N			
7	order_num	순서	int(11)	N			
8	perm_write	등록가능여부	ENUM('Y','N')	N		'Y','N'	
9	perm_edit	수정가능여부	ENUM('Y','N')	N		'Y','N'	
10	perm_delete	삭제가능여부	ENUM('Y','N')	N		'Y','N'	
11	ins_dt	등록일	datetime	N		YYYY-MM-DD HH:MI:SS	
12	edit_dt	수정일	datetime	Y		YYYY-MM-DD HH:MI:SS	

한글테이블명		코드B		영문테이블명		TC_CODEB	
NO	컬럼ID	컬럼명	타입	NULL	KEY	비고	
1	idx	idx	int(11)	N	PK		
2	cd_a	코드A	varchar(10)	N			
3	cd_b	코드B	varchar(8)	N			
4	cd_nm	코드B명	varchar(255)	N			
5	cd_memo	메모	varchar(255)	Y			
6	cd_memo2	메모2	varchar(255)	Y			
7	del_flag	숨김여부	char(1)	N			
8	order_num	순서	int(11)	N			
9	ins_dt	등록일	datetime	N		YYYY-MM-DD HH:MI:SS	
10	edit_dt	수정일	datetime	Y		YYYY-MM-DD HH:MI:SS	

구분		획득(수집) 단계	정제 단계	가공(라벨링) 단계
데이터 구분		원시데이터	원천데이터	최종데이터
데이터 형태	여행로그	여행로그 앱 → MariaDB	MariaDB → CSV	CSV 파일 형태로 산출
	관광사진	여행로그 앱 → Cloud	Cloud → 비식별화작업	JPG 파일 형태로 산출
데이터 형태	여행로그	MariaDB	CSV	CSV
	관광사진	JPG	JPG	JPG

#### 4. 데이터 구성

데이터구분	데이터	데이터 명	수량
[3-005-277] 수도권	여행로그 데이터	tc_codea_코드A.csv	각csv 파일별 4,000Set 구성
		tc_codeb_코드B.csv	
		tc_sgg_시군구코드.csv	
		tn_activity_consume_his_활동소비내역_A.csv	
		tn_activity_his_활동내역_A.csv	
		tn_adv_consume_his_사전소비내역_A.csv	
		tn_companion_info_동반자정보_A.csv	
		tn_lodge_consume_his_숙박소비내역_A.csv	
		tn_move_his_이동내역_A.csv	
		tn_mvnmn_consume_his_이동수단소비내역_A.csv	
		tn_tour_photo_관광사진_A.csv	
		tn_traveller_master_여행객_Master_A.csv	
		tn_travel_여행_A.csv	
		tn_visit_area_info_방문지정보_A.csv	
		tn_poi_master_POIMaster.csv	
	gps_Data	n_gps_coord_*.csv [ * = 여행객 ID ]	4,000개
	photo	여행객ID + 순번. jpg	135,183개
[3-005-278] 동부권	여행로그 데이터	tc_codea_코드A.csv	각csv 파일별 4,000Set 구성
		tc_codeb_코드B.csv	
		tc_sgg_시군구코드.csv	
		tn_activity_consume_his_활동소비내역_B.csv	
		tn_activity_his_활동내역_B.csv	
		tn_adv_consume_his_사전소비내역_B.csv	
		tn_companion_info_동반자정보_B.csv	
		tn_lodge_consume_his_숙박소비내역_B.csv	
		tn_move_his_이동내역_B.csv	
		tn_mvnmn_consume_his_이동수단소비내역_B.csv	
		tn_tour_photo_관광사진_B.csv	
		tn_traveller_master_여행객_Master_B.csv	
		tn_travel_여행_B.csv	
		tn_visit_area_info_방문지정보_B.csv	
		tn_poi_master_POIMaster.csv	
	gps_Data	n_gps_coord_*.csv [ * = 여행객 ID ]	4,000개
	photo	여행객ID + 순번. jpg	160,237개

[3-005-279] 서부권	여행로그 데이터	tc_codea_코드A.csv	각csv 파일별 4,000Set 구성
		tc_codeb_코드B.csv	
		tc_sgg_시군구코드.csv	
		tn_activity_consume_his_활동소비내역_C.csv	
		tn_activity_his_활동내역_C.csv	
		tn_adv_consume_his_사전소비내역_C.csv	
		tn_companion_info_동반자정보_C.csv	
		tn_lodge_consume_his_숙박소비내역_C.csv	
		tn_move_his_이동내역_C.csv	
		tn_mvnmn_consume_his_이동수단소비내역C.csv	
		tn_tour_photo_관광사진_C.csv	
		tn_traveller_master_여행객_Master_C.csv	
		tn_travel_여행_C.csv	
		tn_visit_area_info_방문지정보_C.csv	
		tn_poi_master_POIMaster.csv	
	gps_Data	n_gps_coord_*.csv [ * = 여행객 ID ]	4,000개
	photo	여행객ID + 순번. jpg	161,444개
[3-005-280] 제주도 및 도사지역	여행로그 데이터	tc_codea_코드A.csv	각csv 파일별 4,000Set 구성
		tc_codeb_코드B.csv	
		tc_sgg_시군구코드.csv	
		tn_activity_consume_his_활동소비내역_D.csv	
		tn_activity_his_활동내역_D.csv	
		tn_adv_consume_his_사전소비내역_D.csv	
		tn_companion_info_동반자정보_D.csv	
		tn_lodge_consume_his_숙박소비내역_D.csv	
		tn_move_his_이동내역_D.csv	
		tn_mvnmn_consume_his_이동수단소비내역_D.csv	
		tn_tour_photo_관광사진_D.csv	
		tn_traveller_master_여행객_Master_D.csv	
		tn_travel_여행_D.csv	
		tn_visit_area_info_방문지정보_D.csv	
		tn_poi_master_POIMaster.csv	
	gps_Data	n_gps_coord_*.csv [ * = 여행객 ID ]	4,000개
	photo	여행객ID + 순번. jpg	269,658개

## 폴더 구조

<ul style="list-style-type: none"> <li>&gt; 277.국내 여행로그 데이터(수도권)</li> <li>&gt; 278.국내 여행로그 데이터(동부권)</li> <li>&gt; 279.국내 여행로그 데이터(서부권)</li> <li>&gt; 280.국내 여행로그 데이터(제주도 및 도서지역)</li> </ul>	<ul style="list-style-type: none"> <li>&gt; 01.데이터</li> <li>&gt; 02.저작도구</li> <li>&gt; 03.시모델</li> <li>&gt; 04.교육자료</li> <li>&gt; 05.초기데이터</li> <li>&gt; 06.품질검증</li> </ul>
<ul style="list-style-type: none"> <li>277.국내 여행로그 데이터(수도권) <ul style="list-style-type: none"> <li>01.데이터 <ul style="list-style-type: none"> <li>수도권 <ul style="list-style-type: none"> <li>CSV <ul style="list-style-type: none"> <li>tc_codea_코드A.csv</li> <li>tc_codeb_코드B.csv</li> <li>tc_sgg_시군구코드.csv</li> <li>tn_activity_consume_his_활동소비내역_A.csv</li> <li>tn_activity_his_활동내역_A.csv</li> <li>tn_adv_consume_his_사전소비내역_A.csv</li> <li>tn_companion_info_동반자정보_A.csv</li> <li>tn_lodge_consume_his_숙박소비내역_A.csv</li> <li>tn_move_his_이동내역_A.csv</li> <li>tn_mvmm_consume_his_이동수단소비내역_A.csv</li> <li>tn_poi_master_POIMaster.csv</li> <li>tn_tour_photo_관광사진_A.csv</li> <li>tn_travel_여행_A.csv</li> <li>tn_traveller_master_여행객 Master_A.csv</li> <li>tn_visit_area_info_방문지정보_A.csv</li> </ul> </li> <li>gps_Data <ul style="list-style-type: none"> <li>tn_gps_coord_a_a000011.csv</li> <li>tn_gps_coord_a_a000012.csv</li> <li>tn_gps_coord_a_a000013.csv</li> <li>tn_gps_coord_a_a000014.csv</li> <li>tn_gps_coord_a_a000016.csv</li> <li>tn_gps_coord_a_a000018.csv</li> <li>tn_gps_coord_a_a000019.csv</li> <li>tn_gps_coord_a_a000021.csv</li> <li>tn_gps_coord_a_a000024.csv</li> <li>tn_gps_coord_a_a000028.csv</li> <li>tn_gps_coord_a_a000029.csv</li> <li>tn_gps_coord_a_a000030.csv</li> </ul> </li> <li>photo <ul style="list-style-type: none"> <li>a00001101002p0001.jpg</li> <li>a00001101099p0001.jpg</li> <li>a00001101099p0003.jpg</li> <li>a00001101099p0007.jpg</li> <li>a00001102002p0003.jpg</li> <li>a00001102002p0004.jpg</li> <li>a00001102002p0005.jpg</li> <li>a00001102002p0006.jpg</li> <li>a00001102002p0007.jpg</li> <li>a00001102002p0008.jpg</li> </ul> </li> </ul> </li> </ul> </li> </ul> </li></ul>	여행로그 데이터
	GPS 데이터
	관광사진 데이터

## 5. 데이터 통계

### 5.1 데이터 구축 규모

구분		구축실적
[3-005-277] 수도권	여행자 정보 (여행자 패널 데이터)	4,000 SET
	동선 정보 (GPS 데이터)	4,000 SET
	활동정보 (여행기록 데이터)	4,000 SET
	활동정보 (여행지 사진 데이터)	135,183 장
	소비 내역 (소비내역 데이터)	4,000 SET
	POI 데이터	1 Set
[3-005-278] 동부권	여행자 정보 (여행자 패널 데이터)	4,000 SET
	동선 정보 (GPS 데이터)	4,000 SET
	활동정보 (여행기록 데이터)	4,000 SET
	활동정보 (여행지 사진 데이터)	160,237 장
	소비 내역 (소비내역 데이터)	4,000 SET
	POI 데이터	1 Set
[3-005-279] 서부권	여행자 정보 (여행자 패널 데이터)	4,000 SET
	동선 정보 (GPS 데이터)	4,000 SET
	활동정보 (여행기록 데이터)	4,000 SET
	활동정보 (여행지 사진 데이터)	161,444 장
	소비 내역 (소비내역 데이터)	4,000 SET
	POI 데이터	1 Set
[3-005-280] 제주도 및 도서지역	여행자 정보 (여행자 패널 데이터)	4,000 SET
	동선 정보 (GPS 데이터)	4,000 SET
	활동정보 (여행기록 데이터)	4,000 SET
	활동정보 (여행지 사진 데이터)	269,658 장
	소비 내역 (소비내역 데이터)	4,000 SET
	POI 데이터	1 Set

### 5.2 데이터 분포

(명/권역별 4,000명)

		수도권		동부권		서부권		제주/도서	
성별	남	1,639	41%	1,563	39%	1,524	38%	1,488	37%
	여	2,361	59%	2,437	61%	2,476	62%	2,512	63%
연령별	20대	1,383	35%	1,335	33%	1,382	35%	1,398	35%
	30대	1,421	36%	1,321	33%	1,376	34%	1,366	34%
	40대	633	16%	652	16%	613	15%	627	16%
	50대 ↑	563	14%	692	17%	629	16%	609	15%
여행 기간별	당일	2,192	55%	1,792	45%	1,895	47%	536	13%
	1박 2일	1,252	31%	1,532	38%	1,551	39%	926	23%
	2박 3일 이상	556	14%	676	17%	554	14%	2,538	63%



## 6. 원시데이터 특성

### 6.1 대상분류

- 여행객 데이터
  - 여행객 인구학적 정보 데이터
  - 동반자정보 데이터
- 이동내역 데이터
  - 여행객의 이동 동선 GPS 데이터
  - 여행 중 일정시간을 체류하게 된 이동내역 데이터
- 방문지 데이터
  - 방문지정보 데이터
  - 방문지 촬영사진(관광사진) 데이터
- 소비내역 데이터
  - 사전소비 내역 데이터 : 여행 전 소비내역
  - 이동경비 데이터 : 여행 중 이동경비 내역
  - 숙박 소비내역 데이터 : 여행 중 숙박의 경우 소비내역 데이터
  - 활동 소비내역데이터 : 여행 중 발생하는 소비내역 데이터

### 6.2 제약조건

- 개인정보 보호법에 의한 정보 제공 동의
  - 모집된 여행객들은 사전설문조사와 교육 절차를 진행한 후 계약을 체결함
  - 계약서에는 개인정보 수집/이용/제공 동의서를 포함하여, 여행객들이 입력하는 정보를 수집하는 데에 대한 동의 절차를 진행함
  - 저작물에 대한 권리 사항을 통해 여행객이 올린 데이터를 AI-Hub에 공개하는 점, 지적재산권을 이용할 권리의 허락, AI-Hub 회원의 이용 허락 등을 명시함

### 6.3 속성

데이터 분류	데이터 항목	테이블 명
여행객	동반자정보	tn_companion_info
	여행객 Master	tn_traveller_master
활동내역	GPS	n_gps_coord_* [ * = 여행객 ID ]

	이동내역	tn_move_his
방문지	관광사진	tn_tour_photo
	활동내역	tn_activity_his
	여행	tn_travel
	방문지정보	tn_visit_area_info
	관광사진	여행객ID + 순번. jpg
소비내역	활동소비내역	tn_activity_consume_his
	사전소비내역	tn_adv_consume_his
	이동수단소비내역	tn_mvnm_consume_his
	숙박소비내역	tn_lodge_consume_his
POI Master	POI Master	tn_poi_master_POIMaster

## 7. 기타 정보

### 7.1 포괄성

- 권역별 4,000명의 데이터를 수집하여 전국 16,000명의 여행로그 데이터를 수집

수도권	서울	경인	계
	1,375	2,625	4,000

동부권	강원	대구	경북	부산	울산	경남	계
	1,561	218	961	558	171	531	4,000

서부권	대전	충남/세종	충북	광주	전남	전북	계
	338	1,109	636	142	963	812	4,000

제주도 및 도서지역	제주	도서지역	계
	3,094	906	4,000

### 7.2 독립성

- 개인정보 보호법에 의한 민감정보는 여행객 Travel ID를 사용하여 관리
  - 여행객 ID를 사용하여 데이터베이스를 확장(새로운 레코드 타입이나 데이터 항목 추가)하거나 데이터베이스를 축소(기존의 레코드 타입이나 데이터 항목 삭제)하면서 개념 스키마를 변경할 수 있으며, 후자의 경우 남아있는 데이터만 참조하는 외부 스키마들이 영향이 없음
- 구축된 DBMS는 개념 스키마가 변경되어도 외부 스키마에는 외부스키마에 영향을 미치지 않으면서 개념스키마에 적용되는 제약조건들을 변경가능
- 하나의 논리적 구조로부터 여러 가지의 상이한 물리적 구조를 지원하여 내부 스키마가 변경되어도 외부/개념 스키마가 영향을 받지 않도록 지원

### 7.3 유의사항

- 법·제도 준수
  - 데이터 획득 대상, 획득방법이 법·제도를 저촉하거나 또는 사회 윤리에 어긋나지 않도록 함
  - 개인정보 및 사생활 보호가 필요한 항목 획득 시, 개인정보보호법 등에 따라 적절한 법적, 기술적 절차를 거친 데이터를 활용하며, 그렇지 않은 데이터는 정제 과정에서 처리될 수 있도록 함
- ※ 법적 절차 : 개인정보 활용 동의, 초상 활용 동의, 명예훼손 가능성 여부 검토 등
- ※ 기술적 절차 : 데이터 유형별로 적용할 수 있는 익명처리 기법 적용
  - ① 수치형 데이터 : 데이터 범주화 등

- ② 텍스트 데이터 : 이름, 민감정보 키워드 데이터 변환 등

- ③ 이미지·동영상 데이터 : 모자이크·블러처리, 크롭(자르기) 등

- ④ 음성 데이터 : 크롭(자르기) 등

- 데이터가 3자 제공 및 대중에 개방에 문제가 없도록 법적요건 및동의서 내용 등을 검토 저작권 보호 대상인 데이터 획득 시 법에 저촉되지 않는 범위 내에서 획득할 수 있는 방안을 마련하며, 저작권 보호 대상 저작물 활용 필요 시 가급적 동의서, 계약서 등을 활용한 서면 자료 확보를 권장

※ 예) 방송국 동영상을 활용할 경우 특정 방송국 이름(로고) 노출 가능 여부

※ 예) 이미지·동영상 내 특정 기업의 로고, 제품 형태 등 노출 가능 여부

- 개인정보활용동의서 및 저작물 활용 동의서 등 법적 요건을 준수하기 위한 관리방안을 마련

### 7.4 관련 연구

- 여행로그를 활용한 고속 하이브리드 여행 상품 추천시스템
- 인플루언서 동반여행의 여행상품 선택속성과 인플루언서 특성이 소비자 만족도 및 반응에 미치는 영향
- 여행 블로그의 품질이 이용자 만족도와 재방문 의도에 미치는 영향
- 관광객의 라이프스타일에 따른 여행상품 선택행동에 관한 연구
- 여행상품유형에 따른 인터넷 여행상품 구매결정요인 중요도 차이에 관한 연구
- 목표지향적 행동모델을 적용한 국내 패키지여행상품 이용자의 이용의도에 관한 연구

## 제2장 데이터 구축

### 1. 데이터 구축 개요

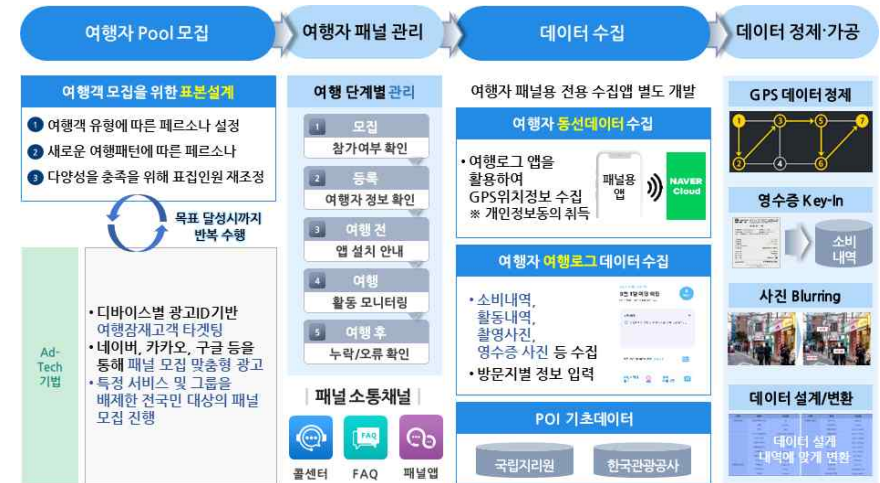
#### 1.1 추진 배경

- 코로나19 이후 관광산업의 디지털 전환 가속화
  - 방역으로 인해 관광객 이동이 제한되면서 새로운 관광 시장을 발굴하는 계기
  - 비대면 서비스에 대한 요구가 높아지면서 관광산업 전용 ICT 기술의 활용이 활성화
- 빅데이터와 인공지능 ICT 기술의 활용에 대한 수요 증가
  - 관광분야에 챗봇, 예약, 모빌리티 등 빅데이터와 인공지능 기술 니즈 증대
- 초개인 맞춤형 여행 및 관광 서비스 본격화 돌입
  - 2000년대 전후 패키지 여행에서 개별 여행 변화 이후 개인 맞춤형 여행 확산
  - 여행 준비 절차의 편의를 높여 개인에게 최적화된 여행 경험 제공
- 1인 여행 및 1인 세대 비중 증가
  - 1인 가구 포함 소규모 가족 비중은 지속적 증가, 3인 이상 가구의 비중은 감소
  - 2021년 관광부문 소비액 중 1인 가구 관광소비 비중 약 14.6% (전년대비 5.5% 급증)
- 전 세계 상위 4개 OTA 그룹 온라인 여행시장 약 97% 과점
  - 익스피디아, 부킹홀딩스, 트립닷컴, 에어비앤비 4개 글로벌 기업의 점유율 비중 심화
  - 세계 관광산업에서 온라인 유통채널의 비중이 2025년에는 72% 예상

#### 1.2 데이터 구축 전략

- 국내 여행로그 데이터 활용 목적을 정의 하고 활용 목적을 기반으로 추진전략을 도출
  - 페르소나별 여행자모집으로 다양한 데이터를 구축

### 1.3 데이터 구축 프로세스



활용목적 정의		<ul style="list-style-type: none"> <li>여행자들의 이동패턴, 소비패턴, 활동패턴을 분석하기 위한 데이터 구축</li> <li>여행자들을 대상으로 관광 개인화 추천 서비스를 개발하는데 필요한 데이터 구축</li> <li>여행자들의 소비 패턴을 분석하여 고부가가치 여행상품을 개발 할 수 있는 데이터 구축</li> </ul>
추진전략 도출		<ul style="list-style-type: none"> <li>최신 선진기법(Ad-Tech 등)으로 여행자모집</li> <li>여행조사 전문기업의 여행자패널 관리</li> <li>공간 전문기업에 의해 GPS/POI 데이터를 정제</li> <li>정형데이터에 대한 정합성 검증 진행</li> </ul>
사업 내용	데이터 구축	<ul style="list-style-type: none"> <li>여행자 Pool 모집: 여행자 표본 설계 (성/연령/소득/숙박/이동수단 최신 선진기법 활용, Ad-Tech 방식의 Digital Marketing Platform)</li> <li>여행자 패널 관리: 여행조사 전문기업의 관리 (문체부 여행조사 전문기업 다양한 패널 소통채널 운영, 콜센터 운영, 패널앱 공지/질의응답)</li> <li>데이터 수집: 여행자용 패널앱 개발 (목표 수량 충족 시까지 데이터 수집)</li> <li>데이터 정제/가공: GPS/POI 정제, 영수증 Key-In, 개인정보 비식별화, 데이터 설계/변환</li> <li>품질검증: 전공정 검증, 전수 검사, 교차 검사, TTA 수검</li> </ul>
	시모델 개발	<ul style="list-style-type: none"> <li>1차년도: 여행자 정보 기반 고지출 여행객 예측 모델, 여행자 선호도 기반 여행강조 추천모델</li> <li>2차년도: 여행 추천 서비스 개발 (안)</li> </ul>

## 2. 임무 정의

### 2.1 임무 정의

- 관광업계 자체적으로 수집하기 어려운 양질의 AI데이터 제공
  - 숙박, 여행업, 요식업 등에서 데이터 분석을 통한 매출 증대 및 활로 개척 시도
  - 가장 수요가 큰 관광객 동선 데이터와 소비내역, 활동내역 등을 학습용 데이터로 제공
- AI기술을 활용한 관광산업 혁신 생태계 구축
  - 관광산업을 혁신하기 위해서는 AI 등의 새로운 디지털기술 적극 도입 필요
  - 고객행동 기반의 분석이 가능하도록 고객들의 동선과 소비 등의 활동 데이터를 구축
- AI기술 기반의 개인화된 서비스로 관광객들의 경험 향상
  - AI를 통해 관광객의 행동을 이해하여 관광객에게 개인화된 서비스를 제공
  - AI를 기반으로 개인화된 맞춤형 관광 서비스를 제공함으로써 관광객의 경험을 향상

### 2.2 데이터 구축 유의사항

- 개인 정보 보호 문제
  - 본 과제는 그 목적 상 개인의 거주지를 비롯하여 관광사진에서 개인 얼굴 이미지 및 자동차 번호판 등 그에 따른 개인 정보가 포함되어 수집하게 되어 개인 정보 보호와 초상권 문제에 직면하게 된다. 여행객 대상자뿐만 아니라 본 과제의 가공과 감수에 참여할 과제 참여자들도 데이터를 취급함에 있어 개인 정보 보호와 기밀 유지 문제와 관련된 위험에 노출되어 있음
- 대처 방안
  - 관광사진의 촬영시 본 과제의 목적, 개요, 개인 정보 보호 정책에 대한 충분히 설명
  - 촬영 대상자는 개인정보보호법에 의거하여 법률전문가들이 검토하고 작성한 '개인 정보 이용 동의서'에 동의 후 서명
  - 가공 및 감사에 참여하는 과제 참여자들은 법률전문가들이 검토하고 작성한 '기밀유지 계약서'에 동의한 후 서명한다.
  - ISO27001에 의거하여 개인 정보를 포함한 자료들을 체계적으로 관리한다.
  - 데이터 수집용 시스템(PC포함)의 사용이력을 관리하고 DB를 암호화하여 사용한다

## 3. 획득(수집)

### 3.1 원시데이터 선정

- 디지털마케팅 플랫폼 기반 여행객 모집
  - Ad-Tech 기술인 DMP(Data Management Platform)을 통해서 온라인상의 각종 활동 기록을 토대로 잠재 여행 고객을 알아내고, 이를 이용해 광고를 하는 방법
  - 개인 비식별 디바이스 키값인 ADID, Cookie 등을 기준으로 사용자들의 온라인 행태 정보를 분석
  - 이러한 기법을 통해서 표본에 최대한 맞는 패널 Pool을 모집

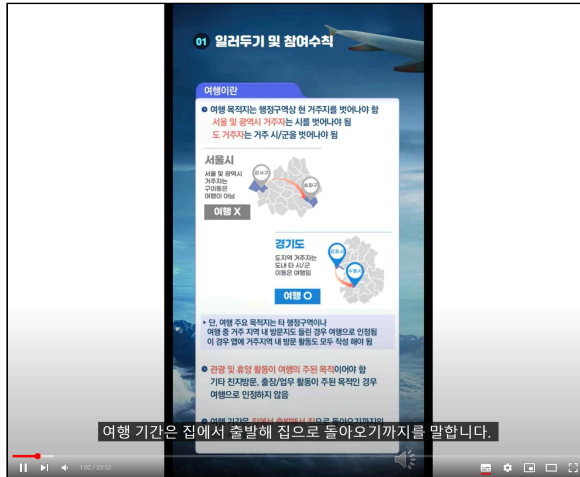


- 디지털마케팅 화면

A형	B형	C형
다양한 여행일정에 대한 안내 마케팅	여행자에 대한 어필형 마케팅	당일여행 집중 모집을 위한 구체적 내용 마케팅

### 3.2 획득(수집) 절차

- 여행객 사전설문 및 교육
  - 여행객들로부터 사전설문을 진행하여, 여행객들의 성별/연령/소득수준 등의 데모그래픽 정보와, 여행일정, 방문지 등의 대략적인 여행 관련 정보를 수집함
  - 또한 여행로그 앱 사용에 대한 교육동영상을 필수적으로 시청하도록 함
  - 사전설문과 교육에 응하지 않은 여행객들은 탈락되며, 이는 추후 데이터 입력을 성실하게 할 여행객들을 선별하는 절차이기도 함
  - 교육 동영상 내용으로 퀴즈를 진행하여 동영상 시청 여부를 확인함



[여행객 대상 교육 동영상]

- 여행로그 앱 계정 부여 및 설치 확인
  - 여행로그 앱 계정은 교육 완료, 사전 조사 완료한 여행객에 한해 제공하며, 계정은 여행객별로 개별 아이디 및 비밀번호 생성
  - 여행로그 앱 설치 및 로그인 여부는 시스템으로 확인하며, 로그인하지 않는 경우 상담원의 전화 독려를 진행
  - 최초 로그인 시 GPS 메타정보 등으로 실제 로그인 여부를 확인
- 여행 및 기록 이행 확인
  - 여행 시작 2일 전 알림 문자를 전송
  - 여행 당일 여행을 시작하였는지, GPS 데이터 기록이 저장되는지도 함께 확인하여, 문제가 있는 경우 여행객에게 연락
  - 여행 종료 3일 이후에도 여행 기록을 시작하지 않는 경우, 여행 상세 기록을 할 것을 안내하며 상세 기록을 모두 완료할 때까지 모니터링
  - 여행객이 작성 완료한 여행 기록에 대해 시간, 장소, 활동 등의 타당성을 검토하며,

이 과정에서 관광지 사진, 영수증 사진 등을 함께 확인

- 관광 활동이나, 소비 활동이 없는 경우, 여행지가 거주지와 동일 시군인 경우 등 여행 기록 내용에 대한 보완이나, 폐기 사항이 있는지 확인
- 보완 및 폐기 사항이 있는 여행 기록에 대하여 여행객이 수정·보완할 수 있도록 안내하며, 보완 및 수정이 완료된 경우 여행 지원금 지급 대상으로 전환

여행자 스마트폰의 패널용 앱을 활용하여 활동기록/사진 등의 데이터를 수집



### 3.3 획득(수집) 기준

- 여행 형태에 따른 페르소나 정의



<여행 형태별 페르소나 선정 및 인원 배분>

- 기존 국민여행조사 데이터를 바탕으로 여행 형태에 영향이 큰 변수를 선정
- 여행유형별 규모를 확인한 후 각 권역별로 여행자수가 많은 Segment를 선정하여 이를 페르소나로 선정
- 권역별로 할당된 여행자 수를 페르소나별로 배분하고 미션을 배분

권역	여행 형태별 페르소나
수도권	서울 여행 수도권 거주 39대 이하 나홀로 여행
	서울 여행 수도권 거주 39대 이하 커플 여행
	서울 여행 수도권 거주 39대 이하 3인 이상 친구 여행
	서울 여행 수도권 거주 40대 이상 3인 이상 친구 여행
	서울 여행 수도권 외 거주 39대 이하 나홀로 여행
	서울 여행 수도권 외 거주 39대 이하 커플 여행

권역	여행 행태별 페르소나
	서울 여행 수도권 외 거주 39대 이하 3인 이상 친구
	서울 여행 기타
	경인 여행 수도권 거주 39대 이하 나홀로 여행
	경인 여행 수도권 거주 40대 이상 나홀로 여행
	경인 여행 수도권 거주 39대 이하 커플 여행
	경인 여행 수도권 거주 39대 이하 3인 이상 친구 여행
	경인 여행 수도권 거주 40대 이상 3인 이상 친구 여행
	경인 여행 수도권 거주 40대 이상 자녀 동반 여행
	경인 여행 수도권 거주 39대 이하 부부 여행
	경인 여행 수도권 거주 40대 이상 부부 여행
	경인 여행 수도권 외 거주 39대 이하 나홀로 여행
	경인 여행 수도권 외 거주 39대 이하 커플 여행
	경인 여행 기타
	강원 여행 수도권/강원 거주 39대 이하 나홀로 여행
동부권	강원 여행 수도권/강원 거주 40대 이상 나홀로 여행
	강원 여행 수도권/강원 거주 39대 이하 커플 여행
	강원 여행 수도권/강원 거주 39대 이하 3인 이상 친구 여행
	강원 여행 수도권/강원 거주 40대 이상 3인 이상 친구 여행
	강원 여행 수도권/강원 거주 40대 이상 자녀 동반 여행
	강원 여행 수도권/강원 거주 39대 이하 부부 여행
	강원 여행 수도권/강원 거주 40대 이상 부부 여행
	강원 여행 기타
	대구/경북 여행 경상권 거주 39대 이하 나홀로 여행
	대구/경북 여행 경상권 거주 39대 이하 커플 여행
	대구/경북 여행 경상권 거주 39대 이하 3인 이상 친구 여행
	대구/경북 여행 경상권 거주 40대 이상 3인 이상 친구 여행
	대구/경북 여행 경상권 거주 39대 이하 나홀로 여행
	대구/경북 여행 경상권 외 거주 39대 이하 커플 여행
	대구/경북 여행 경상권 외 거주 39대 이하 3인 이상 친구 여행
	대구/경북 여행 경상권 외 거주 40대 이상 3인 이상 친구 여행
	대구/경북 여행 기타
	부산/울산/경남 여행 경상권 거주 39대 이하 나홀로 여행
	부산/울산/경남 여행 경상권 거주 39대 이하 커플 여행
	부산/울산/경남 여행 경상권 거주 39대 이하 3인 이상 친구 여행
	부산/울산/경남 여행 경상권 거주 40대 이상 3인 이상 친구 여행
	부산/울산/경남 여행 경상권 거주 39대 이하 나홀로 여행
	부산/울산/경남 여행 경상권 외 거주 39대 이하 커플 여행
	부산/울산/경남 여행 경상권 외 거주 39대 이하 3인 이상 친구 여행
	부산/울산/경남 여행 기타
서부권	충청 여행 충청권 거주 39대 이하 나홀로 여행
	충청 여행 충청권 거주 39대 이하 커플 여행
	충청 여행 충청권 거주 39대 이하 3인 이상 친구 여행

권역	여행 행태별 페르소나
	충청 여행 충청권 거주 40대 이상 3인 이상 친구 여행
	충청 여행 충청권 거주 39대 이하 부부 여행
	충청 여행 충청권 외 거주 39대 이하 나홀로 여행
	충청 여행 충청권 외 거주 40대 이상 나홀로 여행
	충청 여행 충청권 외 거주 39대 이하 커플 여행
	충청 여행 충청권 외 거주 39대 이하 3인 이상 친구 여행
	충청 여행 충청권 외 거주 40대 이상 3인 이상 친구 여행
	충청 여행 충청권 외 거주 40대 이상 자녀 동반 여행
	충청 여행 충청권 외 거주 40대 이상 부부 여행
	충청 여행 충청권 외 거주 39대 이하 부부 여행
	충청 여행 기타
	호남 여행 호남권 거주 39대 이하 나홀로 여행
	호남 여행 호남권 거주 39대 이하 커플 여행
	호남 여행 호남권 거주 39대 이하 3인 이상 친구 여행
	호남 여행 호남권 외 거주 39대 이하 나홀로 여행
	호남 여행 호남권 외 거주 40대 이상 나홀로 여행
	호남 여행 호남권 외 거주 39대 이하 커플 여행
	호남 여행 호남권 외 거주 39대 이하 3인 이상 친구 여행
	호남 여행 호남권 외 거주 40대 이상 3인 이상 친구 여행
	호남 여행 호남권 외 거주 39대 이하 부부 여행
제주 및 도서지역	호남 여행 호남권 외 거주 40대 이상 부부 여행
	호남 여행 기타
	제주 및 도서 지역 여행 39세 이하 나홀로 여행
	제주 및 도서 지역 여행 40세 이상 나홀로 여행
	제주 및 도서 지역 여행 39세 이하 커플여행
	제주 및 도서 지역 여행 40세 이상 커플여행
	제주 및 도서 지역 여행 39세 이하 3인 이상 친구 여행
	제주 및 도서 지역 여행 40세 이상 3인 이상 친구 여행
	제주 및 도서 지역 여행 40세 이상 자녀 동반 여행
	제주 및 도서 지역 여행 39세 이하 부부 여행
	제주 및 도서 지역 여행 40세 이하 부부 여행
	제주 및 도서 지역 여행 기타

- 여행 테마별 페르소나 정의
  - 여행 테마별 페르소나는 최근 이슈가 되고 있는 여행이나 기존의 다른 데이터에서 수집이 어려운 대상을 선정
  - 본 과업에서 선정한 테마별 페르소나는 아래와 같음

여행 테마별 페르소나	설명
Well-ness 여행	웰빙(well-being) + 행복(happiness) + 건강(fitness), 정신적·사회적인 안정과 신체적인 건강의 조화를 이루는 목적의 여행
SNS 인생샷 여행	여행 경험을 소셜미디어에 남기는 목적의 여행



여행 테마별 페르소나	설명
호캉스 여행	고급 호텔 및 리조트를 중심으로 주변 지역 여행
신규 여행지 발굴 여행	유명 관광지나 사람이 많은 곳이 아닌 새로운 관광지를 찾아 나서는 여행
반려동물 동반	반려동물과 함께하는 여행
인플루언서 따라하기 여행	여행 관련 유명 유튜버의 여행 동선에 따른 여행
친환경 여행	조깅 및 쓰레기 줍기를 동반한 플로깅 여행 등 친환경 여행
등반 여행	산을 등반하고 인근 관광지를 방문하는 여행

### 3.4 획득(수집) 조직

[ 여행자 선정 시 여행자와 관리자 역할 및 선정 과정 ]

대상	흐름	설명
여행자	페르소나 선택	○ 페르소나 유형 중 여행 컨셉에 맞게 선택
	미션 선택	○ 여행 중 미션에 대해 선택
관리자	여행계획 검토	○ 선택한 페르소나 및 미션에 부합되는 검토 ○ 방문 시군구 및 방문지별 누적 빈도 확인
	추가 여행지 및 활동 유도	○ 유명 방문지만 계획된 경우 추가 방문 여부 질문
여행자	여행 계획 보완	○ 부적합 요소에 대한 피드백 및 보완 요청
관리자	여행자 선정	○ 페르소나 및 미션을 충분히 반영하고, 방문지 및 여행활동이 다양한 여행자를 최종 선정

#### ○ 콜센터 운영

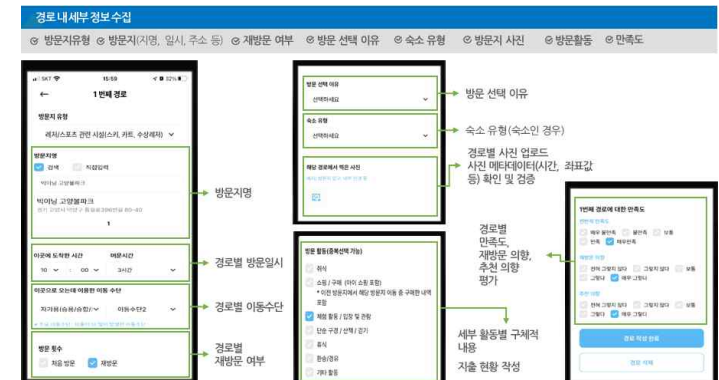
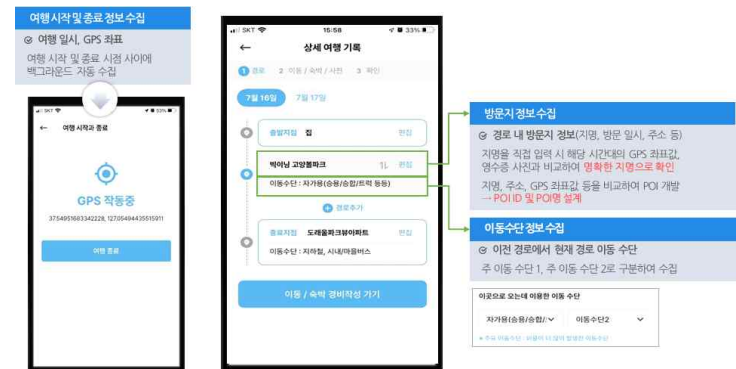
- 여행객들의 문의사항은 전화 또는 카카오톡 채널을 통해 진행
- 콜센터 운영요원들은 클라우드워커들로 구성하였으며, 사전 교육과 대응매뉴얼 등을 통해 일관된 질의응답을 할 수 있도록 함



[콜센터 현장 및 카카오톡 상담채널 운영화면]

### 3.5 획득(수집) 도구

- 데이터 입력을 위한 앱 개발
- 여행객들로부터 데이터를 입력받기 위한 스마트폰용 앱 개발

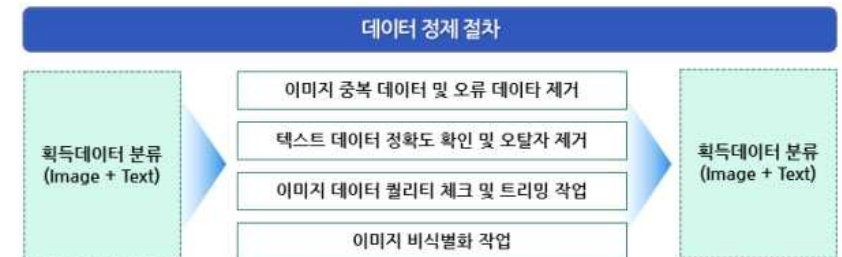


## 4. 정제

### 4.1 원천데이터 규모

구분	데이터명	데이터 설명	포맷	데이터
권역별 구축	여행동선	출발지, 목적지, 경유지를 포함한 GPS 이동 데이터 · 위치(X,Y), 일시(분), 상태(출발,이동중,도착,정차)	CSV	권역별 4,000세트
	소비내역	여행 출발부터 종료까지 소비한 모든 지출금액 · 일시(분), 상호, 주소, 금액, 결제형태(카드,현금) ※ 여행 1세트당 최소 입력건수 10회	CSV	권역별 4,000세트
	활동기록	여행지 주된 활동에 대하여 앱을 통해 입력한 정보 · 여행지명, 방문시간, 만족도(평점), 여행 선호유형, 동행자유무, 이용교통수단, 숙박 타입, 관광지 특성	CSV	권역별 4,000세트
	여행지 사진	여행활동 중 직접 촬영한 사진 데이터 · 개인/민감 정보 비식별화	JPG	권역별 80,000장
	여행자 프로필	개인정보를 제외한 여행자 기본 정보 · 성별, 연령, 거주지, 여행목적, 소득수준	CSV	권역별 4,000세트
구분	데이터명	데이터 설명	포맷	데이터
공통 구축	POI	국가관심지역정보 내 6개 업종(관광, 레저, 예술, 이벤트, 숙박, 음식)에 대한 수도권 내 POI정보	CSV	1백만건
	통계분석	표준공통정보와 여행로그 데이터 간 매칭을 통하여 빈도분석, 상관관계분석, 회귀분석 등에 활용	CSV	1식

### 4.2 정제 절차



[ 원시데이터 정제 절차 ]

- 획득한 원시데이터(Raw Data)에 대하여 본 과제에서 구축하는 데이터셋의 목적에 필요한 형식으로 맞추거나 불필요한 중복을 제거하며, 개인정보를 비식별화하여 처리하는 등 일련의 전처리 과정을 통해 원천데이터(Source Data) 확보
- 어노테이션 단계에 들어가기 전 학습용 데이터에 적합한 데이터를 선별하고 처리하는 정제 프로세스를 수집 방법별로 수행
- 국내 여행로그 데이터는 '이미지' 데이터와 '텍스트' 데이터로 구성되며, 데이터 정제 작업도 '이미지' 데이터 정제와 '텍스트' 데이터 정제로 구분하여 작업

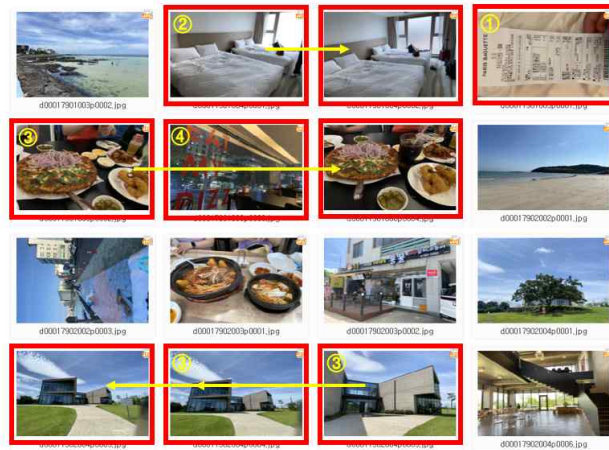


- AI 학습용 데이터의 정확도는 대부분 수집데이터의 품질에 좌우되는 만큼 고품 질의 원천데이터를 확보할 수 있도록 필요한 작업을 적용하여 정제 작업을 수행
- 유사경로 중복성 처리 방안
  - GPS경로 데이터를 분석하여 서로 다른 여행자가 같은 시기에 같은 경로로 여행한 경우 우 악의적인 사례에 대해서는 중복 처리하여 1건만 적용
  - 중복에 대한 판단은 동선과 시간이 동일한 경우를 추출하여 케이스별로 검사

### 4.3 정제 기준

- 촬영사진 데이터 정제
  - 여행로그 앱을 통해 여행객들이 업로드한 촬영사진들에 대한 정제 작업을 진행
  - 사람 위주의 촬영, 초점 불량, 중복 이미지 등 촬영사진 업로드 기준에 부합하지 않은 데이터들을 선별
  - 4개 권역 총 1,131,571장 이미지 중 선별작업 진행 후 726,522장 작업대상으로 분류
  - 정제가 완료된 사진 데이터는 블러링 가공 작업 단계로 이관

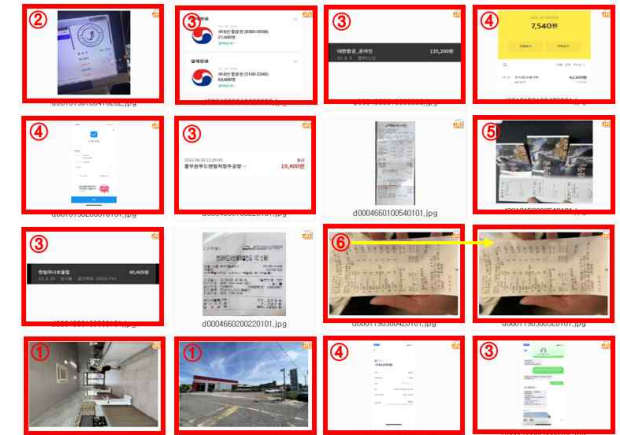
- 촬영사진 정제단계  
- 삭제대상으로 분류
- ① 영수증 사진
  - ② 중복사진
  - ③ 중복으로 판단
  - ④ 의미없는 사진



[ 관광 촬영사진 선별 정제 작업 ]

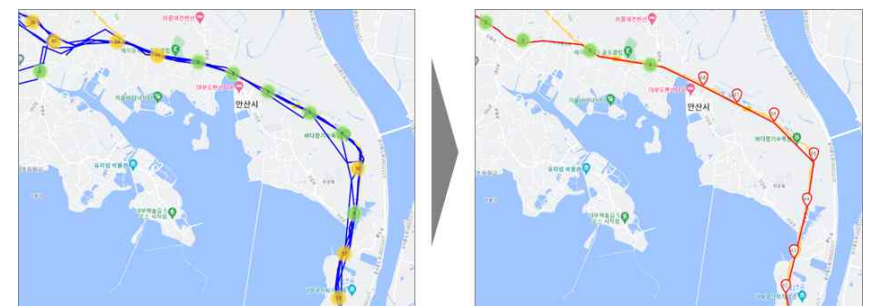
- 영수증 데이터 정제
  - 소비내역은 영수증을 사진으로 찍어 앱을 통해 업로드하도록 하였으며, 정제단계에서 기준에 맞지 않은 사진들을 선별함
  - 흐릿하며 판별이 어려운 사진, 영수증이 아닌 사진, 정보가 너무 부족한 사진, 중복 이미지 등을 선별하여 제외
  - 4개 권역 총 136,281장 이미지 중 선별작업 진행 후 119,049장 작업대상으로 분류
  - 정제가 완료된 영수증 데이터는 Key-In 가공 단계로 이관

- 영수증사진 정제단계  
- 삭제대상으로 분류
- ① 촬영사진(배경)
  - ② 포스트화면
- 일부분만 입력
- ③ 문자메세지
  - ④ 이체내역
  - ⑤ 입장권/승차권
  - ⑥ 중복사진



[ 영수증 사진 선별 정제 작업 ]

- POI 정제
  - 국토지리정보원의 국가기본도, 지명DB와 외부 원천정보 수집처의 공공 행정정보 수집하여 POI 원천데이터로 활용
  - 여행객들이 입력한 방문지 정보와 POI 정보를 맵핑하여, 표준에 맞지 않게 입력된 방문지명은 POI 원천데이터를 기준으로 수정
  - POI 원천데이터에 존재하지 않은 방문지명은 보완 입력
  - 이러한 정제작업을 통해 총 8,135,479 건의 POI 마스터 데이터 구축
- GPS 정제
  - 패널용 여행로그 앱의 GPS 모듈을 통해 여행객들의 동선데이터를 수집하며 이에 대한 정제 작업을 진행
  - 스마트폰 GPS 오차로 인해 실제 위치와 다르게 튜는 좌표들을 정제하여 삭제하거나 원래 지점으로 수렴시키는 작업을 진행



[GPS 정제 전과 정제 후]

#### 4.4 정제 조직



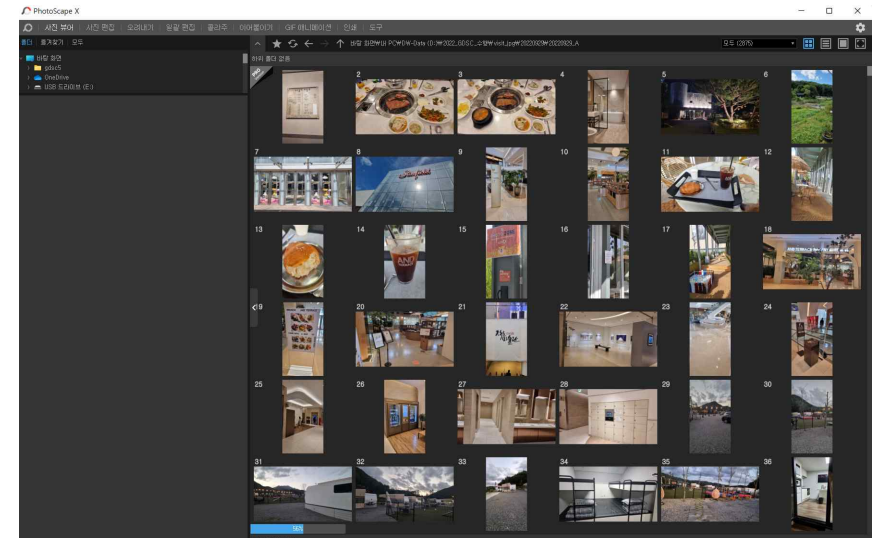
[ 관리자 업무 체계화 ]

##### ○ 교육훈련 진행

교육과정	교육내용	교육일정	교육대상	비고
기본교육	- 기본적인 오리엔테이션 - 인공지능 학습용 데이터 설명	채용 시	클라우드워커, 작업 참여자	
수집/정제 교육	수집/정제 설명 작업 가이드 및 도구 사용법 작업 시연	작업 시작 시, 필요시	클라우드워커, 작업 참여자	
데이터 가공 (라벨링작업) 교육	- 라벨링작업 설명 작업 가이드 및 도구 사용법 시연 및 잘못된 예/ 잘못된 예	작업 시작 시, 필요시	클라우드워커, 작업 참여자	
데이터 검수 교육	- 검수 기준 및 체크 포인트 설명 작업 가이드 및 도구 사용법 시연 및 오류 유형 교육	작업 시작 시, 필요시	클라우드워커, 작업 참여자	

#### 4.5 정제 도구

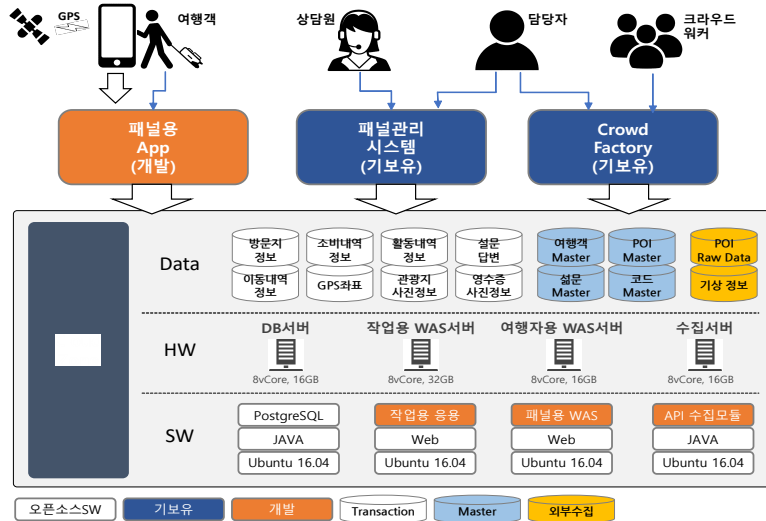
- 이미지 뷰어 및 이미지 편집기능을 갖춘 프리웨어 PhotoScapeX를 활용



[ PhotoScapeX 이미지 확인 ]

## 5. 가공

### 5.1 가공 절차



### 5.2 가공 기준

- 촬영사진 블러링 작업
  - 촬영사진에 대해서는 개인정보 비식별화를 위해 사람 얼굴과 자동차 번호판에 대해 블러링 작업을 진행
  - 블러링 작업은 PhotoScape라는 프리웨어 소프트웨어를 활용
  - 클라우드워커들은 할당된 이미지를 다운로드받아 PhotoScape SW를 활용하여 작업을 진행



[ 비식별화 처리 작업 ]

### ○ 영수증사진 데이터 Key-IN 작업

- 방문지에서의 소비내역은 여행객들이 영수증을 사진으로 찍어 앱을 통해 업로드하도록 되어 있음
- 영수증이 없는 경우에는 여행객들이 앱을 이용하여 직접 입력하는 경우도 있으며, 최종적으로는 하나의 DB 테이블에 저장이 됨
- 영수증 사진을 입력한 경우에는 클라우드워커들이 직접 해당 내용을 Key-In 작업을 통해 DB에 입력
- 금액, 소비항목, 상호, 사업자등록번호, 주소 등을 입력하며, 영수증 유형에 따라 빠진 항목들이 발생할 수 있으나, 최대한 정보를 수집한다는 차원에서 입력 가능한 항목들은 모두 입력하도록 함
- 영수증 Key-In 작업은 컨소시엄이 보유한 클라우드워커 플랫폼을 통해서 입력하며, 해당 작업에 특화하여 별도 기능을 개발

The screenshot shows the '소비내역 데이터 관리' (Consumption History Data Management) interface. It contains the following fields and values:

- 여행객 ID: TR\_000012
- 여행객 명: 홍길동
- POI ID: P\_0002135
- POI 명: 황금어장
- 소비 유형: 식사
- 소비 유형: 신용카드
- 결제일시: 2020-01-25 11:47
- 결제금액: 68,000 원
- 소비 항목: 갈치구이, 전복비빔밥, 공기밥추가, 사이드
- 만족도: ★★★★★
- Buttons: 취소, 저장

[ 영수증 소비내역 입력 데이터 확인 ]

### 5.3 가공 조직

- (가공데이터 관리자) 데이터 라벨링 완료 후 관리 기본 사항
  - 목적에 맞는 데이터 어노테이션 기준을 수립하고 데이터 사용 목적에 맞게 관리함
  - 데이터의 어노테이션 항목은 기존의 데이터 사용 목적의 변화로 인해 수정이 불가피한 경우를 제외하고는 쉽게 바뀌지 않음
  - 데이터의 어노테이션 정보는 쉽게 이해할 수 있어야 하며, 의미가 불분명하여 발생하는 혼란을 최소화함
  - 데이터의 사용 목적에 맞는 일관된 자료인지 수시로 확인함
  - 데이터들의 편향성을 확인 후 필요에 따라 데이터를 지속적으로 추가함





## 6. 검사

### 6.1 검사 절차

#### ○ 촬영사진 데이터 블러링 검사 절차

- 블러링 작업된 데이터를 교차검수를 통하여 반려/승인 처리
- 인물의 경우 작거나 흐릿하여 식별이 어려운 경우에도 블러링되어 있지 않으면 오류로 처리함
- 뒷모습이나 옆모습으로 얼굴이 잘 나오지 않은 경우에도 블러링되어 있지 않으면 오류로 처리함

■ 인물: 블러링 처리는 했으나, 흐림으로 작업된 데이터 오류 리스트 작성



■ 차량: 누락된 차량번호 오류 리스트 작성



[블러링 검사 후 오류리스트 작성]

#### ○ 영수증사진 데이터 Key-IN 검사 절차

- 영수증 Key-In 작업을 진행한 CrowdFactory 프로그램을 동일하게 사용하여 검수를 진행하고 오류가 발생한 경우 반려 처리함
- 특히, 소비내역의 경우 오타가 많이 발생하기 때문에 중점적으로 검수를 진행

#### ○ 의미정확성 검사 절차

- 데이터 검증기관과 검증항목을 함께 도출하여 상호 협의하고, 해당 기준으로 검수를 진행함
- sql로 검증이 가능한 부분은 자동 검수를 진행하였고, 일부 항목들은 육안으로 전수 검수를 진행

[여행객 및 소비내역 로그 정확성 진단규칙]

검사항목	관련 테이블/필드	진단규칙
여행동기와 동반자 수 부합하는지	여행객마스터.여행동기1,2,3 vs 여행객마스터.여행동반자수	여행동기가 동반자와 상반있는 경우(3또는9) 동반자수가 1미만이면 오류
	여행객마스터.여행동반자수 vs 동반자정보	여행객마스터의 동반자 수와 동반자 정보의 count가 다르면 오류

검사항목	관련 테이블/필드	진단규칙
소비유형(활동/이동/숙소)과 소비내역이 일치하는가?	이동수단소비내역	이동수단소비내역에 음식이 있는 경우 이동수단소비내역에 숙박이 있는 경우 이동수단소비내역에 입장권이 있는 경우
	활동소비내역	활동소비내역에 숙박 내용이 있는 경우 활동소비내역에 통행료, 기차, 항공료 등이 있는 경우
	숙박소비내역	숙박소비내역에 음식이 있는 경우 숙박소비내역에 주차비, 통행료, 기차, 항공료 등이 있는 경우 숙박소비내역에 입장권이 있는 경우
	활동유형과 소비내역이 일치하는가?	활동유형과 입력된 소비내역이 불일치 할 경우 오류
	활동내역.입장료구분 vs 활동소비내역.활동유형코드	입장료구분 필드가 null이 아닌데 활동유형이 (체험 활동 / 입장 및 관람)이 아니면 오류
	활동내역 지출구분에 따른 활동소비내역 유무	지출구분이 없음(5)이 아닌데, 활동소비내역에 해당 내역이 없으면 오류 지출구분이 없음(5)인데 활동소비내역에 해당 내역이 있으면 오류

[여행장소 로그 정확성 진단규칙]

검사항목	관련 테이블/필드	진단규칙
방문시작일자가 방문종료일자보다 늦은지?	방문지정보.방문시작일자 vs 방문지정보.방문종료일자	방문시작일자가 방문종료일자보다 늦으면 오류
방문지 방문일자가 여행기간 내에 있는가	방문지정보.방문시작일자, 종료일자 vs 여행.여행시작일자, 종료일자 방문지정보.방문시작일자, 종료일자 vs 여행.여행시작일자, 종료일자	방문시작일자가 여행 시작일자와 종료일자 사이에 있지 않으면 오류 방문종료일자가 여행 시작일자와 종료일자 사이에 있지 않으면 오류
소비결제시각이 방문시작일과 방문종료일 사이에 있는가?	활동소비내역.결제일시_분 vs 방문지정보.방문시작일자, 종료일자	활동소비내역 결제일시가 방문지 시작일자와 종료일자 밖에 있으면 오류
	이동수단소비내역.결제일시_분, 이용 경비구분 vs 여행.여행시작일자, 종료일자	이용경비구분이 주유비, 주차비, 통행료일 때 이동수단소비내역 결제일시가 여행 시작일자와 종료일자 사이에 없으면 오류 이동수단소비내역 결제일시가 여행 종료일자 보다 나중이면 오류
	숙박소비내역.결제일시_분 vs 여행.여행시작일자, 종료일자	숙박소비내역 결제일시가 여행 종료일자 보다 나중이면 오류
해당 방문지역에 불가한 이동방법 입력되었는지?	방문지정보.방문지명 vs 이동수단소비내역.이동수단구분명, 여행객Master.거주지시군구코드	주거지가 제주도가 아닌 여행객의 제주도여행 중 이동수단소비내역에 항공권, 배가 없는 경우
방문지유형과 활동유형이 일치하는가?	방문지정보.방문지유형코드 vs 활동소비내역.활동유형코드	방문지유형이 식당/카페일때 활동소비유형이 취식 또는 휴식이 아니면 오류
		방문지유형이 집, 친구/친지집, 사무실 일때 활동소비유형이 휴식, 취식(배달), 기타활동, 없음 이 아니면 오류

## 6.2 검사 기준

### ○ 적합성[기준적합성]

분류	단계	검사 기준 체크리스트
다양성	구축 계획수립	인공지능이 처리해야 하는 실제 세상의 데이터와 유사한 특성이 데이터에 반영되도록 계획을 수립하였는가?
		인공지능이 처리해야 하는 실제 세상의 데이터와 유사한 변동성을 데이터가 갖도록 계획을 수립하였는가?
	데이터 수집	인공지능이 처리해야 하는 실제 세상의 데이터와 유사한 특성이 데이터에 반영되었는가?
		인공지능이 처리해야 하는 실제 세상의 데이터와 유사한 변동성을 데이터가 갖고 있는가?
신뢰성	구축 계획수립	데이터 수집 시 수집 출처의 객관성 확보를 위한 계획을 수립하였는가?
	데이터 수집	데이터 수집을 위한 수집 출처에 대한 객관성 확보를 위한 근거를 제시하였는가?
충분성	구축 계획수립	인공지능 학습모델에 필요한 분류체계 및 분류체계별 데이터 수집 최소 수량 결정을 위한 절차를 마련하였는가?
		인공지능 학습모델에 필요한 분류체계 및 분류체계별 데이터 수집 최소 수량 결정에 대한 근거를 제시하였는가?
	데이터 수집	인공지능 학습모델에 필요한 분류체계 및 분류체계별 데이터 수집 최소 수량을 확보하였는가?
균일성	구축 계획수립	인공지능 학습모델에 필요한 분류체계별 데이터 수집 수량에 대한 적합한 비율 결정을 위한 절차를 마련하였는가?
		인공지능 학습모델에 필요한 분류체계별 데이터 수집 수량에 대한 적합한 비율 결정을 위한 절차를 마련하였는가?
	데이터 수집	인공지능 학습모델에 필요한 분류체계별 데이터 수집 수량에 대한 적합한 비율에 맞게 수량을 확보하였는가?
사실성	구축 계획수립	인공지능 학습용 데이터 구축 시 데이터 수집이 인위적인 환경이 경우 실제 환경 및 상황의 특성을 반영하기 위한 계획을 수립하였는가?
		인공지능 학습용 데이터 구축 시 데이터 수집이 인위적인 환경인 경우 실제 환경 및 조건이 일관성을 갖도록 계획을 수립하였는가?
		데이터 수집 시 다중 데이터 간에 동기화를 하도록 계획을 수립하였는가?
	데이터 수집	데이터 수집이 인위적인 환경인 경우 실제 환경 및 상황의 특성이 반영된 근거를 제시하였는가?
		데이터 수집이 인위적인 환경인 경우 실제 환경 및 조건이 일관성이 확보된 근거를 제시하였는가? 데이터 수집 시 다중 데이터 간에 동기화된 근거를 제시하였는가?
공평성	구축 계획수립	인공지능 학습용 데이터 구축 시 지역적 편견, 사회적 편견, 인종적 편견 등을 방지하기 위한 계획을 수립하였는가?
	데이터 수집	인공지능 학습용 데이터 구축 시 지역적 편견, 사회적 편견, 인종적 편견 등의 방지 결과에 대한 근거를 제시하였는가?

### ○ 적합성[기술적합성]

- 품질목표 달성기준 : 준수율(%) 95% 이상

$$\text{준수율}(\%) = \frac{\text{준수 건수(데이터수)}}{\text{검사총 건수(데이터수)}} \times 100(\%)$$

[ 검사기준 및 내용 ]

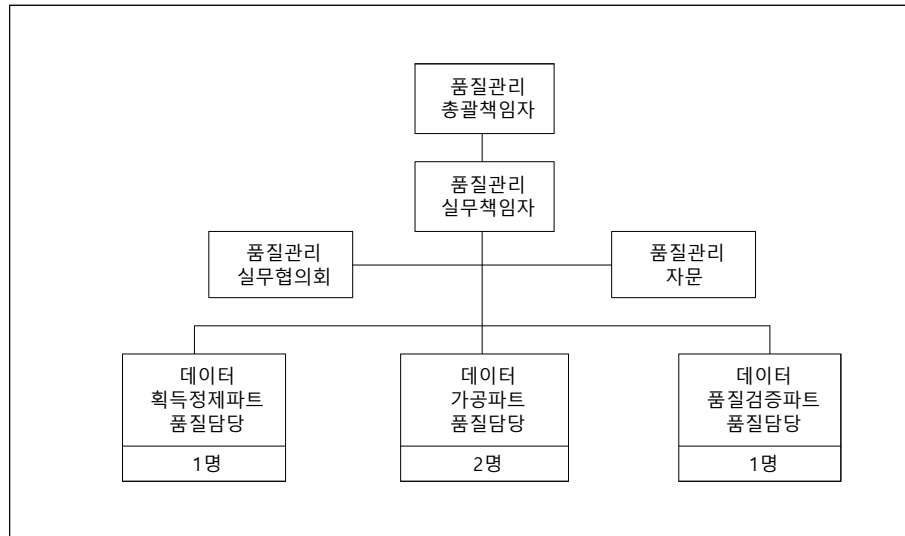
검사대상			검사 방법 (* 전수조사를 기본으로 함)
산출물	데이터 종류	내용	
원시 데이터	<ul style="list-style-type: none"> <li>- 여행자 동선데이터</li> <li>- POI 기초데이터</li> <li>- 여행자 활동내역 데이터</li> <li>- 여행자 소비 내역서</li> <li>- 여행자 촬영 이미지</li> </ul>	값	인공지능 학습용 데이터 구축 시 정의된 파일 포맷을 원시데이터에 적용하고 있는지 전수조사를 통해 확인
			인공지능 학습용 데이터 구축 시 정의된 해상도를 원시데이터에 적용하고 있는지 전수조사를 통해 확인

### ○ 품질 항목 요약표

구분	지표	품질목표
구축 공정	준비성	95% 이상
	완전성	95% 이상
	유용성	95% 이상
데이터 적합성	기준 적합성	95% 이상
	기술 적합성	100%
	통계적 다양성	100%
데이터 정확성	의미 정확성	95% 이상
	구문 정확성	99% 이상

### 6.3 검사 조직

[ 품질관리 조직도 ]



조직 구분	역할과 책임
품질관리 총괄책임자	• 품질 검증의 기준, 수행과정 등을 관리하고 감독하는 역할
품질관리 실무책임자	• 여행로그 학습용 데이터의 품질관리 실무 총괄
품질관리 실무협의회	• 여행로그 학습용 데이터의 참여기관 실무자로 구성하여 품질관리 주요 계획, 실무 이슈 등의 협의
품질관리 자문	• 여행로그 학습용 데이터의 품질관리 주요 계획, 품질 현안 등의 협의
데이터 획득 정제파트 품질관리 담당	• 원시 데이터 및 정제(원천) 데이터의 검수와 블러링데이터 품질관리를 병행하여 진행
데이터 가공파트 품질관리 담당	• 블러링 데이터의 검수와 라벨링 데이터의 품질관리를 병행하여 진행
데이터 품질검증파트 품질관리 담당	• 여행로그 학습용 데이터의 적합성 및 정확성에 대한 품질관리 담당

### 6.4 검사 도구

#### 6.4.1 프로파일링 기법에 의한 검사

- 데이터의 형식을 기준으로 한 검사
- 테이블설계서에 정의된 Not Null, Unique Key 여부, 데이터 타입 등을 기준으로 한 검사를 수행
- 최소/최대 범위가 있는 경우 해당 boundary 내에 있는지 검사

[ 프로파일링 검사 항목 ]

프로파일링 검사 항목	검사 내역	검사 방법
구조 완전성	데이터 구조가 테이블 설계내역에 맞게 되어 있는지 확인	DML 확인
코드 유효성	코드값이 해당 코드테이블에 존재하는지 검사	진단 SQL 실행
형식 유효성	데이터 유형 (Character, Number)과 자릿수가 설계 내역대로 입력되었는지 검사	DML 확인 진단 SQL 실행
여부 유효성	Boolean 데이터의 경우 정해진 규칙에 따라 입력되었는지 확인 (예, Y/N, 0/1 등)	진단 SQL 실행
날짜 유효성	날짜 항목의 경우 해당 형식 준수 여부 확인 (YYYYMMDD or YYYYMMDD hh:mm:ss)	진단 SQL 실행
범위 유효성	숫자 항목의 경우 정해진 범위 내에 있는지 확인	진단 SQL 실행

#### 6.4.2 논리적 오류에 대한 검사

- 데이터의 형식이 아닌 내용의 논리적 오류를 기준으로 검사하며
- 데이터 간의 중복 또는 불량 검사 기준을 마련하고 이를 기준으로 검사
- 진단 SQL을 기반으로 1차 검사하여 오류의심 목록을 만든 후, 하나씩 육안 검사하여 최종 판단

[ 중복/불량 검사 기준 ]

유형	검사 내용	판정 여부
동선 및 활동 데이터	중복 - 같은 방문지가 서로 다른 POI명으로 중복 입력된 경우 (오류 예시: 제주공항 vs 제주국제공항) - 같은 활동내역이 다른 이름으로 중복되어 입력된 경우 (오류 예시: 승마체험 vs 말타기)	중복
	중복 예외 - 같은 방문지가 시간의 차이가 있을 경우 중복 허용 (오류 예시: 여수여행시 낭만포차거리를 방문하고 다음날	적합

유형			검사 내용	판정 여부
	불량		식사를 위해 낭만포차거리를 재차 방문)	
		주소	- 방문지 주소가 여행지역에서 벗어난 경우 (오류 예시: 제주 여행인데 서울 주소)	불량
		시간	- 출발시간이 도착시간보다 늦은 경우 - 방문지와 방문지 도착시간이 물리적으로 불가능한 시간으로 입력된 경우 - 영업시간 외의 시간으로 입력된 경우 (오류 예시 : 전시관 방문시간이 심야 시간으로 입력)	
		이동 방법	- 해당 방문지역에 불가능한 이동방법 입력된 경우 (오류 예시 : 부산 -> 제주, 이동수단(기차) )	
		활동 내역	- 해당 관광지 성격에 맞지 않은 활동내역 입력 (오류 예시 : 해운대 해수욕장 - 등산)	
		방문 목적	- 방문장소와 방문목적이 상식적으로 불일치 (오류 예시 : 박물관 - 번지점프 )	
소비 내역 데이터	중복		- 동일 소비내역이 다른 명칭으로 중복 입력된 경우	중복
	불량	지역	- 방문지역 이외의 지역에서 결제한 영수증	불량
		일시	- 여행기간 외에 결제한 영수증	
		품목	- 여행 목적과 맞지 않은 소비 내역 (오류 예시 : 모바일로 생활용품 구매)	
		금액	- 비이상적인 결제금액 (오류 예시 : 식당에서 1백만원 결제)	

#### 6.4.3 사진 데이터 검사





- 대상 이미지의 구분, 판별이 불가능한 경우 불량으로 판정
- 흔들림, 흐림 : 핀아웃 된 경우 불량
- 노출 과다, 부족 등 : 노출로 인한 대상 판별 불가시 불량
- 방문지가 아닌 곳에서 촬영된 영상또는 여행기간중이 아닌 일시에 촬영된 영상

[ 사진 데이터 검사 기준 ]

분류	유형	기준	이미지 예시
비식별화	사람 얼굴	- 사람 얼굴은 데이터 구축시 비식별화 처리(블러링)함 - 여행자 얼굴 노출 사진 촬영 지양 - 인물이 중앙에 위치한 기념사진류의 사진 촬영 불가	

분류	유형	기준	이미지 예시
		( 관광객 중심의 사진 촬영 ) - 거리/유적지 풍경에서 사람이 포함된 사진은 가능함(추후 공정에서 블러링 또는 모자이크 처리로 비식별화)	
	차량번호판	- 차량번호판은 비식별화 처리함 - 차량을 포함 할 시에는 가급적 차량의 사이드라인 촬영 - 일반차량만 중심으로 촬영된 경우 불가 - 거리풍경에서 일부 차량 포함된 경우 가능함( 추후 공정에서 비식별화 )	
	기타 개인정보	- 거리풍경 사진중 간판은 블러링 처리 함 - 기타 개인정보가 포함된 사진 불가 - 신분증, 신용카드, 금융정보 포함 사진 불가 - 간판이 비중이 전체 사진의 1/4초과 사진 불가	
중복	동일대상 원근 촬영	- 같은 촬영 대상을 멀리서 촬영한 경우 중복 처리하여 2개 이미지중에서 한 장의 사진만 선택함 - 근거리 촬영 영상이 원거리 촬영 영상에 포함된 경우 중복	 (중복으로 택일)
	동일대상 일부 확대	- 같은 촬영대상의 일부를 확대 촬영한 경우 중복 처리 - 중심 이미지가 50%이상 겹치는 경우 중복 처리 - 동일 건물을 찍은 경우라도 앞, 옆, 뒷부분(버드아이) 촬영등은 각각 다른 사진으로 인정 함	 (중복으로 택일)
노출/포커스	흔들림, 흐림 (핀 아웃)	- 포커스가 맞지 않아 핀아웃 된 이미지등 저품질 이미지는 불가 - 사진 자료 구축에서 사진품질을 좌우하는 중요한 문제임)	 ( 적정 )                      ( 핀아웃 )



분류	유형	기준	이미지 예시
	노출 부족	- 노출이 부족으로 전체적으로 어두워 대상 판별이 어려운 이미지 - (예외) 역광으로 촬영한 여행사진으로 대상의 외곽이 선명하고 구분 가능한 경우 가능	  ( 적정 )                      ( 노출부족 )
	노출 과다	- 노출과다로 전체적으로 하이톤으로 뜬 영상	  ( 적정 )                      ( 노출과다 )

#### 6.4.4 사진 데이터 유효성 검수

- 사진데이터의 EXIF 데이터 중 “촬영일자”를 추출하여 여행객이 여행기간에 포함되는지 확인
- 사진데이터의 EXIF 데이터 중 “GPS좌표”를 추출하여 여행동선에 일치하는 지 확인(여행로그 동선데이터와 비교하기 위해 주소 또는 POI데이터로 변환)

#### 6.5 기타 품질관리 활동

- 데이터 검수는 1차 검수팀과 2차 검수팀으로 구성하고, 교차검수를 진행하여 데이터 검사 완료
- 1차 검수에서 불합격 데이터는 가공작업 수행팀에서 재작업 후 검수작업 재진행



## 7. 학습 모델

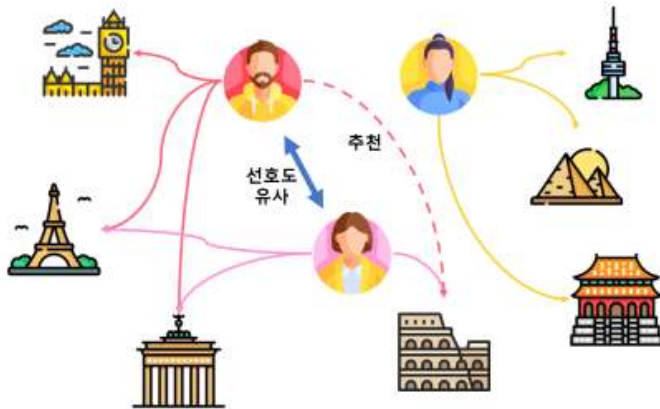
### 7.1 학습 모델 후보

데이터 명	3-005 국내 여행로그 데이터 수집			
학습 모델 후보	알고리즘	성능지표	선정 여부	선정 사유
여행객 정보 기반 고지출 여행객 예측 모델	Pycaret	F1-score 0.70이상	○	- Data Leakage를 막기 위해 여행객의 사전 정보, 페르소나, 소득 수준, 호텔 예약 정보 등 여행 출발 이전부터 알 수 있는 정보를 정제한 후 선정 - EDA를 통해 각 변수별 특성을 파악하고 소비 지출에 영향을 주는 변수에 무엇이 있는지 1차적으로 확인 - 2D Tensor Data의 분류 예측 문제는 일반적으로 트리 기반 부스팅 모형이 가장 좋은 성능을 보이나 데이터 전체 데이터 개수가 크지 않은 경우 Overfitting의 문제를 고려해야 함 - 앙상블 이외의 모형이 더 좋은 성능을 보일 가능성을 배제할 수 없기 때문에 데이터 전처리 이후 Pycaret 알고리즘을 통해 Validation Data Set에 가장 높은 성능을 보이는 모델을 선정하고 추가 작업하여 최종 모델 선정
여행객 선호도기반 여행 장소 추천 알고리즘	Ensemble model	Recall@10 0.25	○	- 추천시스템은 전통적으로 협력필터링, 콘텐츠 기반 시스템 그리고 이 둘의 장점을 합친 하이브리드 모델이 존재 - 최근 인공지능 분야의 비약적인 발전과 더불어 오토인코더와 같은 딥러닝 모델들이 추천시스템에 적용 - 그러나 본 개발 모델의 경우 추천 성능 못지않게 이후 확장 가능성 및 모델 결과에 대한 분석 그 자체가 중요함 - 따라서 기존 추천시스템에서 사용되던 딥러닝 기반 모델보다 사용자 정보를 넣었을 때 선호도를 예측하는 Regression 기반의 모델을 사용하여 추천 장소를 선정하는 방식을 선택 - 일반적으로 2D Tensor 데이터에서 가장 좋은 성능을 보이는 Random Forest, Cat Boost, LGBM, XG Boost 등의 모델을 후보 모델로 선정함

### 7.1.1 여행객 정보 기반 고지출 여행객 예측 알고리즘

- 학습모델 개발 목표
  - 고지출 여행객을 예측하고, 고지출 여행객 예측에 사용된 중요 특성 변수가 무엇인지 파악할 수 있는 인공지능 모델을 개발
- 학습모델 개요
  - 여행지에서 지출을 많이 하는 여행객을 예측하여 분류하는 모델
  - 지출 상위 20% 이상을 '고지출'로 정의하며 총 지출이 threshold 미만인 패널을 0, threshold 이상인 패널을 1로 코딩하여 지출 예측 이진 분류 모델을 학습
  - input 데이터로 여행 이전부터 알 수 있는 여행객의 정보만을 넣어 여행지에서 지출을 많이 하는 여행객이 누구인지 예측
- 학습모델의 활용 방안
  - 여행지에서 지출을 많이 하는 여행객의 후보군을 미리 알 수 있다면 효과적인 여행 마케팅이 가능해져 공공기관, 사기업 모두 홍보 또는 마케팅 대상 고객을 선정하는 일에 있어서 의미 있게 활용 가능
  - 여행지에서 지출이 많은 여행객 집단을 미리 파악하여 고지출 여행객에 맞는 개인별 타겟 마케팅을 통해 마케팅의 효율성 제고
  - 또한 불특정 다수에게 전달되는 마케팅으로 인해 발생하는 비용을 절감하고 여행상품에 관심이 없는 고객이 가질 수 있는 불만을 줄이는 데 도움

### 7.1.2 여행객 선호도 기반 여행 장소 추천 알고리즘



[ 사용자 선호도기반 추천 시스템 개요 ]

- 학습모델 개발 목표
  - 여행객이 선호할만한 구체적 여행 장소를 추천하는 알고리즘 개발
  - 모델 개발 이후 여행객 선호도를 예측하는 알고리즘의 Feature Importance 값을 확인하여 여행객 페르소나별로 어떤 특성 변수가 여행지 선정에 중요한지 개별 특성

### 변수에 대한 분석 진행

- 학습모델 개요
  - 여행지에 대한 정보는 대부분 다른 사람들의 리뷰와 전문가들의 평가에 의해 이뤄지는 경우가 많아, 사용자 개별 특성에 맞는 선호도가 고려되지 않은 추천으로 소비자 만족도 비교적 낮게 형성
  - 자녀 여부, 자차 여부, 어르신 동행 여부, 연령대 등 사용자 개인의 특성을 AI모델에 입력하여 사용자와 비슷한 조건을 가진 사람들이 지금까지 만족했던 방문지 추천
  - 여행지 추천시스템은 추천 성능에 대한 결과만큼 특정 정보를 가진 사용자가 어떤 여행지를 선호하는지 선호도 예측에 사용된 각 특성변수에 대한 분석이 중요
  - 따라서 해당 모델은 일반적인 추천시스템 알고리즘을 사용하지 않고, 딥러닝 모델보다는 분석이 가능한 머신러닝 모델의 Regression 방식을 우선적으로 고려하여 선호도를 예측하는 Regression방식의 추천시스템 알고리즘을 개발
  - 본 개발을 통하여 지금까지 방문지 추천과 달리 사용자가 자신의 특성에 부합하는 개인 맞춤형 방문지를 추천받을 수 있도록 함
- 학습모델의 활용 방안
  - 기존 여행지 추천과 여행 경로 선택은 전적으로 타인의 추천 혹은 리뷰를 바탕으로 이루어졌으나, 대중적인 여행지 선택이 사용자의 취향에 부합할 수 있으나 부합하지 않을 경우 사용자가 불편함을 느낄 수 있음
  - 여행객의 정보를 방문지 추천에 반영하여 사용자의 특성에 부합하는 장소 항목을 추천
  - 또한 각 여행 페르소나별 선호도 예측 모델을 분석하여 어떤 항목이 포함되어 있을 때 어떤 여행지의 선호도 예측 값이 높게 나오는지 확인, 마케팅 및 여행 산업 계획에 포함시킬 수 있음

## 7.2 학습 모델 개발

### 7.2.1 여행객 정보 기반 고지출 예측 모델

- 모델 목적
  - 여행객들의 사전 정보를 토대로 여행 출발 전, 여행지에서 지출을 많이하는 여행객을 미리 예측하여 분류
- 사용 모델 및 선정 이유
  - 본 과제에서 활용할 수 있는 여행객 사전 정보에 범주형 데이터가 다수 존재하여 범주형 데이터를 효율적으로 처리하기 위해 Categorical Boosting Machine(Cat Boost)을 후보군 선정
  - 또한 데이터 특성과 모델 목적상 과적합 최소화를 위해 Cat Boost가 오버피팅을 줄이는 데 이점이 있고 자체적으로 feature importance를 제공하기에 예측 결과에 대한 사후 분석(feature 관련)에 유리하다는 점을 고려
  - 본 과제 1차 샘플 데이터를 활용하여 다른 모델들과 Cat Boost의 성능을 비교했을 때 Cat Boost가 우수한 성능을 보였고, 최종적으로 사용 모델로 선정

- 사용 데이터
  - 여행객 데이터 중 지출 예측에 유의미하다고 판단되는 데이터를 통합하고, data leakage가 없는 데이터를 선별하여 사용
  - 거주지시군구코드, 성별, 최종학력이수여부, 혼인상태, 가족현황, 직업\_기타, 본인소득, 가구소득, 여행빈도\_기간, 여행빈도, 선호여행\_시도(3개), 여행스타일(8개), 여행현황(거주지, 목적지, 동반현황), 여행 동기, 주요이동수단, 여행 페르소나, 사전 소비내역
- 전처리 작업
  - 여행일수 전처리 : 여행 시작 및 종료 날짜 정보를 추출하고 총 여행 일수를 계산
    - 사전 숙박 예약 정보 : 숙박 데이터 결재 정보 중, 여행 일자 이전에 미리 예약한 내역만을 추출하여 '사전 숙소 예약 금액'을 0과 1사이 bin으로 처리
  - 동반자 연령대 : 구체적 연령대를 알 수 없어, 동반자 연령대의 평균을 추출하여 사용
    - feature 내부 데이터 전처리 : 데이터 중 라벨 인코더로 부여할 수 없는 정보들을 추출하고, 각 데이터 형태에 맞게 범주형 변수로 변경  
ex) 3개의 column에 나뉘 있는 범주형 형태의 여행 동기를 binary 형태의 data로 바꿈
- 학습 모델 설계
  - Smote 적용
  - Imbalanced data model pipeline 구축
  - grid search와 random search를 조합한 Hyper parameter 탐색
  - 최종 모델 적합

```
# iteration 1000 (디폴트 세팅)으로 두고, 위에서 구한 best parameter를 적용하여 final model 2를 재작
final_model_2=CatBoostClassifier(subsample=0.7,
max_leaves=31,
max_depth=13,
loss_function='Logloss',
learning_rate=0.02,
boosting_type='Ordered',
cat_features=catgorical_list)

ss=SMOTE(k_neighbors=10)
X_res, Y_res = ss.fit_resample(X_train,y_train)

final_model_2.fit(X_res, Y_res)

500:  learn: 0.0665009  total: 1h 50m 0s  remaining: 1h 54s
983:  learn: 0.0665709  total: 1h 50m 19s  remaining: 1m 47s
984:  learn: 0.0665072  total: 1h 50m 29s  remaining: 1m 40s
985:  learn: 0.0665063  total: 1h 50m 42s  remaining: 1m 34s
986:  learn: 0.0664935  total: 1h 50m 54s  remaining: 1m 27s
987:  learn: 0.0664923  total: 1h 51m 8s  remaining: 1m 21s
988:  learn: 0.0664916  total: 1h 51m 21s  remaining: 1m 14s
989:  learn: 0.0663322  total: 1h 51m 29s  remaining: 1m 7s
990:  learn: 0.0662799  total: 1h 51m 41s  remaining: 1m
991:  learn: 0.0662793  total: 1h 51m 54s  remaining: 54.1s
992:  learn: 0.0662535  total: 1h 52m 5s  remaining: 47.4s
993:  learn: 0.0662256  total: 1h 52m 15s  remaining: 40.7s
994:  learn: 0.0662096  total: 1h 52m 22s  remaining: 33.9s
995:  learn: 0.0661374  total: 1h 52m 34s  remaining: 27.1s
996:  learn: 0.0660965  total: 1h 52m 45s  remaining: 20.4s
997:  learn: 0.0660952  total: 1h 52m 57s  remaining: 13.6s
998:  learn: 0.0659759  total: 1h 53m 9s  remaining: 6.8s
999:  learn: 0.0659505  total: 1h 53m 21s  remaining: 0us

<catboost.core.CatBoostClassifier at 0x13e307e60a0>
```

[고지출 예측모델 학습 스크린샷]

- 성능 평가
  - 이진 분류를 평가하기에 적합한 f1\_score을 기준으로 사용
  - 학습결과는 f1\_score 성능 목표를 넘겼으며 test data set을 활용하여 검증한 결과는

아래와 같음

F1-Score	precision	recall	목표 달성 여부
0.8067	0.8286	0.7829	0

- confusion matrix:

```
In [18]: from sklearn.metrics import confusion_matrix
cf = confusion_matrix(y_test, y_test_pred)
print(cf)

[[2455  104]
 [ 137 503]]
```

## 7.2.1 여행객 선호도 기반 여행 장소 추천 모델

- 모델 목적
  - 유저 정보와 여행지역(시/도) 정보가 주어지면 10개의 여행지를 추천하는 모델
- 사용 모델 및 선정 이유
  - [Cat Boost] 특성변수가 범주형일 때 일반적으로 사용하는 one-hot encoding 대신 실수인 순서목표통계량(ordered target statistic)으로 전환하여 사용하는 방식
  - 범주형 특성변수가 많을 때 적합한 모델임. 본 과제에서는 범주형 데이터인 여행지명, 시/도, 군/구 정보를 학습하기에 용이하기 때문에 CatBoost를 모델로 선정
- 사용 데이터 (사용 변수)
  - (Input data) 유저 정보 및 여행지 정보
  - (output) 추천 여행지 10군데
  - 데이터 중 활용 변수는 유저 정보와 여행지 정보로 구분하여 사용

**유저 정보** : 유저 미션, 성별, 연령대, 소득, 여행스타일(8개 항목), 여행동기(1개 항목), 동반자 수

**여행지 정보** : 여행지명, 여행지 시/도 및 군/구 정보, 여행지 종류, 해당 여행지 체류시간 평균, 추천 의향 점수 평균, 재방문 의향 점수 평균, 재방문 여부 비율, 동반자 수 평균 만족도

- 전처리 작업
  - 여행지 선별 : 방문지 유형코드 중 1-자연관광지, 2-역사/유적/종교 시설 (문화재, 박물관, 촬영지, 절 등), 3-문화시설(공연장, 영화관, 전시관 등), 4-상업지구(거리, 시장, 쇼핑시설), 5-레저/스포츠 관련 시설(스키, 카트, 수상레저), 6-테마시설(놀이공원, 워터파크), 7-산책로, 둘레길 등, 8-지역축제, 행사에 해당하는 방문지를 여행지로 파악해 데이터 사용
  - 시/도 변수, 군/구 변수 생성 : 여행지의 주소에서 시/도와 군/구 변수 생성  
예시) 인천 강화군 삼산면 매음리 629 → 시/도 변수: 인천, 군/구 변수: 강화군
  - 학습데이터에서 여행지에 대한 평균 변수 생성 : 학습데이터에서 각각의 여행지마다 체류시간 평균, 추천의향 점수의 평균, 재방문 비율, 동반자 수의 평균, 재방문의향 점수의 평균을 산출해 변수 생성

- 학습모델 설계
  - Random Search를 활용한 초모수 조절
    - K-Fold Cross Validation를 병행하여 가장 검증된 초모수 값 확보
    - 데이터셋에 고유한 관광지가 많아 학습 시 최대한 많은 데이터를 보존하기 위해 K=10, 10개의 fold로 교차검증 진행
  - CatBoost Regressor를 적용해 만족도 예측
    - 만족도: 1(매우 불만족), 2(불만족), 3(보통), 4(만족), 5(매우 만족)
  - 모델이 예측한 여행지의 만족도가 4.5이상이면 추천 항목에 포함
    - ※ 추천의 기준으로 설정한 4.5는 도메인에 따라 조정이 가능하며 본과제에서는 사용자가 관광지에 대한 만족도를 4점, 5점으로 주는 경향이 높아 보수적으로 4.5를 임계값으로 설정
- 성능 평가
  - Recall@10
 

$$Recall@10 = \frac{10개\ 추천\ 항목\ 중\ 사용자가\ 관심\ 있는\ 아이템\ 의\ 개수}{사용자가\ 관심\ 있는\ 모든\ 아이템\ 의\ 개수}$$

    - ※ 사용자가 만족하는 모든 아이템 중에서 모델이 추천한 아이템 10개가 얼마나 포함되는지 비율을 의미하며 각 사용자마다의 recall@10 값을 구하여 그 평균을 최종 recall@10 값으로 산정
  - 최종 성능 : 0.3745 (목표 성능 이상)

## 제3장 데이터 활용

### 1. 데이터 활용

데이터 명	3-005 국내 여행로그 데이터 수집
학습 모델	여행객 정보 기반 고지출 여행객 예측 모델
모델	Pycaret
성능 지표	F1-score 0.70이상
개발 내용	여행객들의 사전 정보를 토대로 여행 출발 전, 여행지에서 지출을 많이하는 여행객을 미리 예측하여 분류
응용서비스 (예시 및 유의사항)	여행지에서 지출이 많은 여행객 집단을 미리 파악하여 고지출 여행객에 맞는 개인별 타겟 마케팅을 통해 마케팅의 효율성 제고하고, 불특정 다수에게 전달되는 마케팅으로 인해 발생하는 비용을 절감하고 여행상품에 관심이 없는 고객이 가질 수 있는 불만을 줄이는 데 도움
학습 모델	여행객 선호도 기반 여행 장소 추천 모델
모델	Essemble model
성능 지표	Recall@10 0.25
개발 내용	유저 정보와 여행지역(시/도) 정보가 주어진다면 10개의 여행지를 추천하는 모델
응용서비스 (예시 및 유의사항)	여행객의 정보를 방문지 추천에 반영하여 사용자의 특성에 부합하는 장소 항목을 추천하고 각 여행 페르소나별 선호도 예측 모델을 분석하여 어떤 항목이 포함되어 있을 때 어떤 여행지의 선호도 예측 값이 높게 나오는지 확인, 마케팅 및 여행 산업 계획에 포함시킬 수 있음

#### 1.1 관광업계에서 즉각 활용할 수 있는 양질의 AI데이터 제공

- 인공지능은 관광산업의 변화를 주도할 주요 기술 중 하나로 인식되고 있고 (전효재·한희정, 2018; UNWTO 홈페이지1)) 숙박, 여행업, 요식업 등의 다양한 분야에서 데이터 분석을 통한 매출 증대 및 활로 개척을 시도해 왔으나, 개별적으로 양질의 데이터 확보가 어려웠음
- 특히 AI는 자연언어 처리와 기계 학습을 통해 관광객의 행동을 이해할 수 있기 때문에, 관광객에게 개인화된 서비스를 제공함으로써 관광객의 경험을 향상시킬 수 있음(한희정, 한국문화관광연구원 2021) . 코로나 등의 외부변수로부터 많은 영향을 받은 분야이기 때문에 관광객 데이터 분석 기반의 경영계획 수립이 매우 중요함
- 관광객에 대한 고객 클러스터링은 공통적인 욕구와 행위 패턴에 의해 정의된 고객 클러스터에 맞는 인공지능 기반의 개인화 맞춤형 서비스를 제공하기 위해 개인 데이터가 필요함(주신옥, 2018)

- 관광업계에서 가장 수요가 많은 관광객들의 동선 데이터와 소비내역, 활동내역 등을 AI학습을 할 수 있는 형태의 데이터로 제공

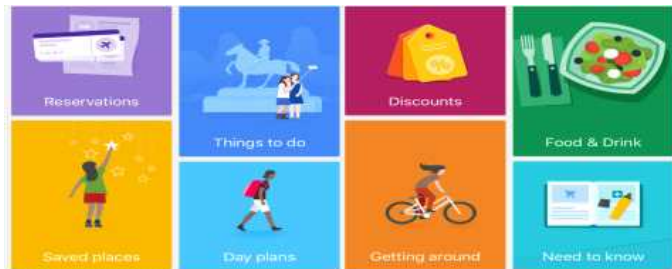
## 1.2 관광산업 혁신 생태계를 구축할 수 있는 데이터와 활용방안 제공

- 관광산업을 혁신하기 위해서는 AI 등의 새로운 디지털 기술들을 적극적으로 도입해야 하며, 이를 토대로 생태계를 구성하는 환경변화와 혁신을 도모
- 고객행동 기반의 분석이 가능하도록 고객들의 동선과 소비 등의 활동 데이터를 구축하고 관광데이터 활용의 선진사례를 제시
- AI를 통해 관광객의 행동을 이해하여, 관광객에게 개인화된 서비스를 제공함으로써 관광객의 경험을 향상
- 여행 시장 활성화 대비 '항공사, 호텔, 크루즈 회사 및 기타 여행업체' 모든 고객 접점에서 마찰을 없애고 고객 여정을 따라 개별적으로 고객과 소통하기 위해 고객 경험을 우선시 제공하도록 챗봇에 활용 : 여행 사용자 중 37%가 여행 계획을 준비할 때 챗봇 이용을 선호 분석

## 2. 응용 서비스

- (실시간 맞춤형 여행서비스) 여행 중 빈번히 변경되는 여행계획에 맞춰 최적화된 행선지 경로 탐색, 관광지 운영시간을 고려한 관람 최적 동선, 당일 소화 가능한 스케줄 최적화 등
  - 구글 여행 플래너 구글트립은 구글이 찾은 방대한 자료와 메일 계정과 연동된 개인정보의 조합으로 막강한 맞춤형 서비스를 제공.
  - 여행 기록, 항공권과 호텔, 레스토랑, 차량 렌트, 익스커션 등 예약 정보를 자동 감지. 다음 여행을 위한 항공권 예약 정보, 도시의 볼거리와 즐길 거리, 할인 정보 등을 제시 외에 방문지 일정에 맞는 루트를 제안하고, 이동 교통수단과 예상 시간, 입장료, 후기 등 제공

<사례 구글트립: Google Trips>



- (AI기반 챗봇) 관광객 문의 사항을 실시간으로 해결해 항공사, 여행사, 관광지에서 활용
  - 빠른 문제 해결과, 개인화된 대응, 최고의 편리함으로 여행객 기대치 제고

- WhatsApp, Facebook Messenger 및 Google Assistant에서 Netomi 기반의 다국어 AI 가상 도우미인 Juliet은 AI 챗봇으로, 인간의 개입 없이 고객 서비스 티켓 중 74%를 자율적으로 해결
- 코로나19 팬데믹 상황에서 서비스 통화가 45배 폭증하여 수 백 가지 질문에 답변할 수 있는 Juliet은 수만 통에 달하는 상담원의 부담 해소
- Juliet을 출시한 이후 고객 만족도(CSAT)가 24% 상승
- (AI기반 안내로봇) 호텔, 박물관 등 관광지에서 AI기반 안내서비스 로봇을 통해 관광객에게 최신화된 관광정보 제공

## 3. 응용서비스 개발

- 모델 공개
  - 공유 플랫폼(Git hub)을 통하여 모델 공유 및 기술 지원
  - 개발 인공지능 모델의 구조 공유
  - 딥러닝 학습이 완료된 모델의 가중치 파일 공유
  - 공개 모델 및 가중치 파일의 사용 방법 공유
  - 본 사업에서 수집한 인공지능 학습 데이터셋의 활용 방법 공유



[ Git-hub를 통한 소스 공개 예 ]

- 모델 공개 방법
  - 서비스 코드 및 활용 예제를 오픈소스 형태로 Git-hub를 통한 공개



[ AI 모델 공개 도식도 ]

#### 4. 기술 지원

- AI-Hub를 통한 온라인 기술지원
  - 여행로그 데이터 AI 개발을 위해서는 관광산업 트렌드와 관련 기술동향 등에 대한 지식이 추가로 필요
  - 여행로그 데이터셋은 다양한 유형의 데이터들의 조합으로 구성되기 때문에 수요자들이 활용할 때 질의사항이 많을 것으로 예상됨
  - 이러한 문제점을 해결하기 위해 온라인 게시판 등을 통해 개발자들의 궁금증을 해결하여 여행로그 데이터를 기반으로 한 AI 개발을 쉽게 할 수 있도록 기술지원을 함
- 데이터셋을 활용한 AI 성능 검증 지원
  - 데이터 수집 시 공개하지 않는 검증 셋을 별도로 마련하고, 해당 데이터를 기반으로 개발이 완료된 AI의 성능 검증을 지원
  - AI 개발자들에게 목표의식을 주고, 회사에서는 특정 문제를 해결할 수 있는 검증된 다수의 AI의 성능을 비교
  - 공정한 성능평가를 통해 개발자가 만든 AI의 객관적 성능을 확인할 수 있도록 함
- AI 개발자 커뮤니티에 참여 및 홍보
  - 주기적으로 꾸준히 교류할 수 있는 AI 개발자, 커뮤니티 및 관련 비즈니스 개발 주체 등에 적극적으로 참여 및 홍보를 통하여 본 과제에서 구축한 데이터셋 활용할 수 있도록 함
  - 구축된 데이터에 대한 피드백 및 수정, 여행로그 데이터의 개발 방향 논의, 전문가, 산업계의 미팅 등을 주선하여 실제로 현장에서 데이터가 어떻게 사용되는지, 데이터와 AI 기술의 융합 개발을 위해 현장의 지식과 현황을 논의