요약 보고서

팀명	DATA 189	
세부 주제(제목)	근교여행을 바탕으로 한 새로운 관광 권역 형성	
분석 데이터	**문화·관광 데이터** (1종 이상 사용 필수)	□ 국민문화예술활동조사 □ 국민여가활동조사 ■ 국민여행조사 □ 주요관광지점입장객통계 □ 근로자휴가조사 □ 외래관광객조사 ■ 신한카드
	기타 활용 데이터	CODE.csv (분석 데이터를 시각화하기 위한 지역 코드명 데이터) TL_SCCO_SIG.shp(http://www.gisdeveloper.co.kr/ 대한민국 최신 행정구역(SHP)) 네이버 블로그 크롤링데이터(naver_여주여행_검색결과.xlsx, naver_춘천여행_검색결과.xlsx, naver_가평여행_검색결과.xlsx, naver_함안여행_검색결과.xlsx, naver_창녕여행_검색결과.xlsx, naver_합천여행_검색결과.xlsx, naver_예산여행_검색결과.xlsx, naver_천안여행_검색결과.xlsx, naver_평택여행_검색결과.xlsx) stopwod_list.xlsx(https://mr-doosun.tistory.com/24 1차 불용어 리스트 데이터) phploeBuh.png(워드클라우드 작성을 위한 도형)
분석 도구	□ SPSS □ SAS ■ R ■ PYTHON □ STATA □ EXCEL □기타	
기획 내용	코로나 19로 인한 생활 속의 변화는 일상이 되었으며 이는 뉴노멀로 자리 잡을 것입니다. 이는 일상뿐만 아니라 다양한 분야에서 영향을 주고 있으며 관광 또한 큰 타격을 입음으로써 여행 트렌드에 대한 변화와 새로운 형태의 관광이 성행할 그것으로 예상합니다. 실제로 코로나 19로 인한 불안감이 증대되고 안전한 여행을 추구하는 사람들이 많아짐에 따라 여행지 선택 시 이동 거리의 중요성이 2019년 대비 2020년 2.3%나 증가하였습니다. 또한, 2021년 여행 트렌드에 따르면 수도권, 대표 관광지에 대한 수요가 감소하고 거리상 가까운 지역으로의 이동이 증가했다는 것을 알 수 있습니다. 따라서 신한카드 데이터를 활용하여 뉴노멀을 정의한 다음 실제 근교 여행의 트렌드 가능성의 근거를 제시하고 국민 여행조사를 바탕으로 관광지의 특징을 정의하였습니다. 이후 분석 결과를 종합하여 근교 여행을 활용한 새로운 기대효과 및 패러다임을 제시하였습니다.	
데이터 활용 분석	〈신한카드 데이터 - 내국인〉 1. 먼저 데이터에서 V1(회원거주지) 변수에 존재하는 13,652개의 결측치를 확인하여 결측치가 존재하는 행을 제거한 후 후 분석 진행하였습니다. 이후 관광에 대한 뉴노멀을 정의하기 위해 “GB3(이하 업종 대분류)” 변수의 값이 “여행”인 경우를 추출하고 “TA_YM(이하 이용 연월)”을 기준으로 정렬하여 “GB2(이하 업종 소분류)”별 “USEC(이하 이용 건수)”의 추이를 시계열 그래프를 통해 나타냈습니다. 코로나 최초 발생 시기부터 이용 건수가 하락하는 경향을 보이지만 코로나 19 1차 대유행(2020년 3월)을 기준으로 점차 회복하는 추세를 보여 2020년 3월 기점으로 이후에 대한 시점을 “뉴노멀”로 정의하였습니다.	

	다음으로 “V2(이하 가맹점 주소)”를 관광지역으로 간주하고 제주 지역 관광지에 대한 회원 거주지 별 이용 건수를 시계열 그래프로 나타낸 결과 다른 지역에 비해 빠른 속도로 회복하는 추이를 확인하였습니다. 이는 코로나 19의 영향으로 해외여행이 불가능해지면서 해외여행의 대체 관광지로 선호하여 나타나는 현상으로 판단하여 분석 목적에 부합하지 않아 회원 거주지, 가맹점 주소가 제주인 경우를 제외하였습니다. 2. 뉴노멀의 기준으로부터 여행_체험(업종 대분류_업종 소분류)에 대한 변화를 확인하기 위해 업종 소분류 변수의 값이 체험인 데이터에 대하여 뉴노멀의 기점인 2020년 3월을 기점으로 뉴노멀 이전, 이후로 나눈 두 개의 최종 데이터를 생성하였습니다. 이때 수도권(서울, 경기, 인천)에 해당하는 값이 크기 때문에 체험 활동에 대한 지역별 이용 건수의 정확한 비교를 위해 최종 데이터에 대해 표준화된 비율을 구했습니다. 뉴노멀 이전, 이후에 대한 각각의 데이터를 회원거주지와 가맹점 주소 변수를 기준으로 이용 건수의 총합과 회원거주지가 동일한 지역에 대한 총합을 계산한 다음 두 값을 나누어 새로운 파생변수를 추가하였습니다. 이후 지역별 이용 건수 비율의 평균을 뉴노멀 이전 이후로 나누어 피벗 테이블 형태로 나타낸 후 차이를 계산하여 뉴노멀 이전 대비 이후에 대한 지역별 이용 건수 평균 증감률을 확인하였습니다. 그 결과 경기, 충남뿐만 아니라 대부분 지역에서 거주지로부터 근거리에 있는 지역에 대한 여행의 비율이 높은 증감률을 보였으며 이를 바탕으로 뉴노멀 시대의 근교 여행의 트랜드 가능성을 확인하였습니다. 〈국민 여행조사〉 1. 먼저 2018, 2019, 2020년 국민 여행조사 데이터를 선택하여 필요한 변수인 “ID, 여행활동, 여행 방문지”를 선택하였습니다. 이후 각 연도에 대한 데이터로부터 “여행경험_” 혹은 “여행횟수_” 변수를 바탕으로 실제 관광객 여부를 판단하여 tour_survey…_tourist(…은 연도를 나타냄) 데이터를 생성하였습니다. 위 데이터를 두 개로 분할 하였는데 melt 함수를 활용하여 “_방문지역” 혹은 “_방문지” 변수를 ID를 기준으로 녹여낸 tour_survey…_melt.dat 데이터 그리고 ID 변수와 “여행 활동”변수를 추출한 tour_survey…_activity 데이터를 만들었습니다. 이때 tour_survey…_activity 데이터의 활동변수는 지역별 각 활동에 대한 빈도수를 계산할 수 있도록 1, 0 코딩을 하여 전처리하였습니다. 이렇게 만든 tour_survey…_melt.dat과 tour_survey…_activity 두 데이터를 ID를 기준으로 결합하였고 unique 함수를 활용하여 같은 ID에 중복되는 방문지역이 있는 경우 제거하였습니다. 이후 aggregate 함수를 활용하여 “방문지역” 변수를 기준으로 방문지역별 각 여행 활동 빈도수를 나타내는 tour…_area_activity 데이터를 생성했습니다. 연도별 세 데이터를 더해 3년간의 지역별 여행 활동 빈도수 데이터 tour_area_activity를 생성하고 각 여행 활동변수에 대한 평균이 10 이하인 변수를 제거하는 과정을 거쳤습니다. 또한, 신한카드 분석 결과 대부분의 광역시와 특별시의 관광 비육이 감소하여 따라 특별시, 특별자치시, 광역시를 제외하였고 제주도의 경우 해외여행 대체지역으로 판단했기 때문에 제외하였습니다. 이후 관광지의 편중으로 각 활동 빈도수에 대한 지역별 차이가 존재하는 것을 확인하여 이를 해결하기 위해 해당 지역의 각 활동 빈도수를 해당 지역의 활동 빈도수 총합으로 나누어 비율로 나타낸 test.dat_proportion 데이터를 생성하여 분석을 진행했습니다. test.dat_proportion 데이터에 대하여 elbow plot을 그린 결과 군집의 개수가 4개 혹은 5

	개일 때 그래프의 기울기가 급격하게 변화하는 지점이라고 판단하였고 이에 따라 군집 개수를 각각 4개와 5개로 지정하여 kmeans 군집분석 진행하였습니다. 이후 실루엣 계수를 확인했으나 이때 실루엣 계수가 매우 낮아 추가적인 데이터 탐색이 필요하다고 판단했습니다. 2. tour_area_activity 데이터에서 활동변수에 대한 평균을 통해 탐색한 결과 “여행 활동” 변수 간 평균의 차이가 크며 이러한 문제가 분석에 영향을 미쳤을 것으로 판단했습니다. 이에 따라 분석변수를 평균에 대한 순위를 기준으로 총 3개의 데이터로 분할 하였고 test.dat_proportin 데이터와 동일한 방법으로 변환하였습니다. 위 3개의 데이터에 대하여 각각 elbow plot을 그린 후 적절한 군집의 개수를 설정한 이후 kmeans 군집분석을 진행하였고 실루엣 계수를 확인했습니다. 결과적으로 모든 데이터에서 군집이 4개인 경우 실루엣 계수가 비교적 높은 값을 갖으며 군집별 산점도를 확인한 결과 군집의 특성을 잘 나타낸다고 판단하였습니다. 추가적으로 평행좌표 그래프를 통해 각각 데이터에 대한 군집의 특성을 확인했습니다. 이렇게 분석한 결과로 얻어진 각 3개의 군집에 대한 데이터를 지역명을 기준으로 결합하였고 그 데이터를 CODE 데이터(지역명과 지역 코드에 대한 정보를 포함)와 지역명을 기준으로 다시 결합하여 map.dat 데이터를 생성하였습니다. 이후 paste 함수를 이용하여 군집 변수들의 값을 결합하여 지역의 특징을 설명할 수 있는 파생변수를 생성하였습니다. 예를 들어 (가) 지역이 첫 번째 변수들을 사용한 군집분석에서 4번 군집, 두 번째 변수들을 사용한 군집 분석에서 B 군집, 세 번째 변수들을 사용한 군집분석에서 d 군집에 속하는 경우 “4-B-d”로 표현할 수 있습니다. 이를 평행좌표 그래프에 대한 정보를 바탕으로 해석한다면 (가) 지역의 특징은 음식 관광, 역사유적지방문, 지역축제/이벤트 참여 활동의 수요가 높은 지역으로 표현할 수 있습니다. 이렇게 파생변수는 최대 64개의 값을 나타낼 수 있으며 본 분석에서는 총 56가지의 값으로 표현되었습니다. 〈크롤링 데이터〉 네이버에서 제공하는 API를 이용하여 군집분석에서 특성이 각각 다른 지역의 중심을 기준으로 인접 지역에 대한 지역명 + 여행 키워드로 네이버 블로그 게시글 크롤링을 진행하였습니다. 이후 모든 문장을 하나로 병합하고, 토큰화하여 불용어 사전으로 불용어 1차 제거하고 정제된 데이터에서 “Konlpy” 형태소 분석기를 통해 구두점, 특수문자 등을 제외한 명사 말뭉치를 생성하였습니다. 명사 말뭉치에서도 불용어가 확인되어 2차 제거하였고 이후 빈도수가 높은 상위 100개 단어를 선정하였습니다. 이를 바탕으로 워드 클라우드로 시각화하였으며 군집분석을 바탕으로 표현한 지역별 특징과 일치하는 것을 확인하였습니다. 이후 기준지역에서 인접 여행지 3곳에 대한 각각의 상위 100개의 단어를 수집하여 빈도순으로 정렬하였고 상위 30개의 단어만을 추출하여 기준지역 + 인접 지역으로 형성된 새로운 관광 권역에 대한 키워드를 확인하였습니다.
결론	1. 지역 간의 특징을 비교할 수 있으며 이때 인접 지역 간의 관광지에 대한 특징이 유사한 경우 새로운 관광 활동을 개발함으로써 차별화된 관광지 개발이 가능합니다. 이는 해당 지역의 관광 경제 활성화뿐만 아니라 주변 지역에 대한 긍정적 경쟁 효과도 기대할 수 있을 것으로 판단됩니다. 2. 거주지역 근교에 있는 관광지의 활동에 대한 키워드 분석을 진행함으로써 추천 시스템을 개발하며 이를 통해 근거리에서 다양한 활동을 할 수 있는 새로운 관광 시스템으로 활용 가

	능할 것으로 기대됩니다. 마지막으로 문화적 특성, 지역적, 환경적 특성에 따른 관광 권역이 아닌 자신이 사는 지역을 기준으로 즉, 사람을 기준으로 유연하게 형성되는 관광 권역 형성이 가능하여 기존에 없던 새로운 관광 패러다임 제시할 수 있을 것으로 판단됩니다.