4차 6기 세미프로젝트

|  |
| --- |
| 따릉이 편도이용 개선방안 |

**2020년 01월 02일**

**서비스 산업 데이터를 활요한 머신러닝 분석 과정(A)**

4조

권용진

신현정

유지수

정재원

**목 차**

**1. 프로젝트 개요**  **1**

1.1 프로젝트 기획 배경 및 목표 1

1.2 구성원 및 역할 2

1.3 프로젝트 추진 일정 3

**2. 프로젝트 결과**  **7**

2.1 데이터 수집 7

2.2 데이터 분석 8

2.3 데이터 분석 결과 9

**3. 기대 효과**  **10**

3.1 향후 개선 사항 10

3.2 기대 효과 11

**4. 개발 후기**  **12**

**1. 프로젝트 개요 [ 따릉이 편도이용 개선 방안 ]**

1.1 프로젝트 기획 배경 및 목표 [ 따릉이: 대중교통으로서의 역할로서 필요 ]

* 출 퇴근 시간대나, 위치적 입지에 따라 따릉이 편도이용이 대중교통의 기능을 일부 분담하고 있다는 것을 확인하여 교통정책에 반영, 환승 정책 등의 도입

1.2 구성원 및 역할

|  |  |  |  |
| --- | --- | --- | --- |
| 이름 | 전공 | 역할 | 구현 부분 |
| 권용진 | 컴퓨터교육과 | 팀원 | 시각화 연구  -구글, 네이버 등의 경로 표시를 프로젝트 목적에 맞는 형태로 이용하는 방법 연구 |
| 신현정 | 전자정보공학부 | 팀원 | 기초 통계량 작업  - 왕복, 편도 구분  - 시간에 따른 이용량 (년, 월, 일, 시 ,분) |
| 유지수 | 보건의료산업학과 | 팀원 | 최초 전처리  -처리 속도 개선을 위한 변수 정리  - 파생 변수 생성 및 불필요 기호 삭제 |
| 정재원 | 경영학, 경영과학 | 팀장 | 전처리 통합 작업  -기초 통계 작업 병합을 통해,  최종 데이터 추출 |

1.3 프로젝트 추진 일정

|  |  |  |  |
| --- | --- | --- | --- |
| 구분 | 기간 | 활동 | 비고 |
| 사전  기획 | ~12/4 | 프로젝트 기획 및 팀 구성 |  |
| ~12/4 | PJT주제 선정, 팀(PM/팀원) 구성 | 4인/팀 |
| 11/23 | 프로젝트 멘토링  [프로젝트 방향 설정 및 현업프로젝트 소개] | 현업 멘토 참여 |
| PJT  수행  /  완료 |  | 프로젝트 수행  추출한 노선을 수치정보를 포함하여 지도에 시각화 | 데이터 전처리 및  탐색적 분석 |
|  | 권용진   * 데이터에서 추가적으로 위치 시각화를 도출해낼 방법 찾기   신현정   * 최종 프로젝트 발표 전까지 데이터에서 추가적으로 얻어낼 수 있는 데이터 시각화 방법 찾기   유지수   * 최종 프로젝트 발표 전까지 데이터 시각화, 예측모델 아이디어 구상 및 실현방법 찾기   정재원   * 데이터처리 효율화, 처리속도 개선, 변수 탐색 및 모델 도출 |  |
| 12/20 | 세미 프로젝트 중간 발표 | 현재 정제된 따릉이 데이터를 이용하여 그래프로 시각화 도출 |
| 1/4 | 팀별 최종 발표 (구축 완료 보고) | 최우수 한 팀 선발  멘토 평가 |

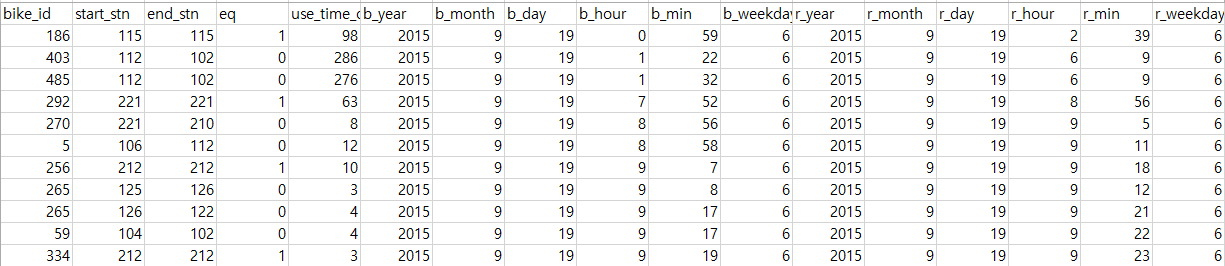
**2. 프로젝트 개발 결과 [따릉이 데이터 전처리 및 시각화]**

2.1 데이터 수집 [따릉이 데이터 수집]

* 이용 데이터: 서울 열린 데이터 광장 -> 서울 공공 자전거 대여 이력
* 2015.09 ~ 2019.11 간의 데이터가 축적 되어 있음
* 자전거 번호, 대여 시간, 대여정류장 번호, 대여정류장 명, 반납 시간, 반납 정류장 번호, 반납 정류장 명, 거치대 번호, 이용시간, 이용거리 등으로 분류되어 있음
* 47여개 파일이 CSV, XLS 등으로 제공 중
* 파일 당 100만 ~ 200만 건 정도의 레코드가 기록되어 있음

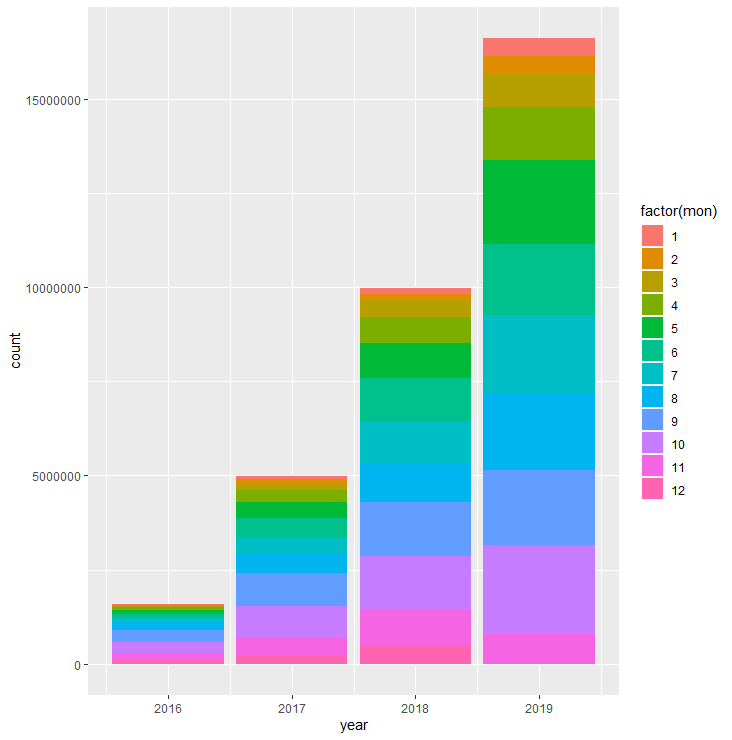
2.2 데이터 분석

* 1차 전처리: 데이터 형태 통일 및 불필요한 요소 제거
* 대여, 반납소 번호: 3,5,7,9996,9997,9998,9999,99999 와 같은 센터 또는 이동정비소로 이동되는 따릉이는 실제 이용하는 따릉이가 아닌 수리되어지는 따릉이 이므로 전처리를 통헤, 불필요 데이터 제거 작업 완료
* “XXXX-XX-XX”, “XX: XX : XX” 와 같은 형태로 년월일, 시분초가 하나의 형태로 기입되어 있었는데, 데이터 전처리 작업 시 편리하도록 각각 년,월,일,시,분,초 분리하여 열 별로 새로운 변수 생성 완료
* 2차 전처리: 분석에 필요한 형태로 가공



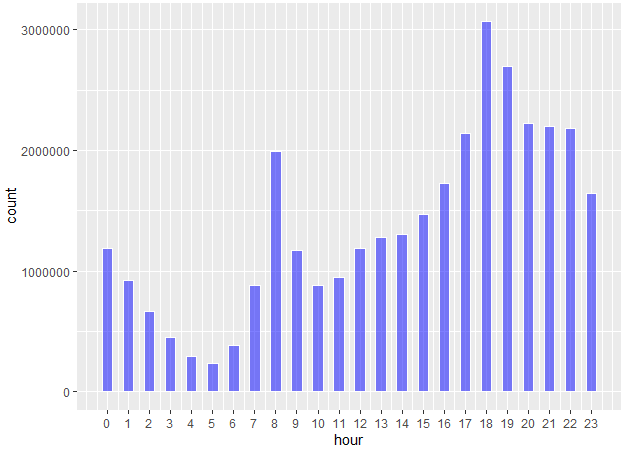
* 기존 데이터에서, 필요한 데이터만 추출하여 분석에 이용하기 편하도록 전처리 완료
  + 정류소 일치 변수 생성: 대여 정류소와 반납 정류소 일치할 경우 1, 다를 경우 0
  + 기존에 기록 되어 있던 이용시간이 0 또는 수치 오류로 기입된 경우가 많아, 반납과 대여 시간 차이를 계산하여 이용시간 계산 변수를 생성
  + 요일 정보 추가하여 저장
  + 처리속도 향상을 위해 사용하지 않을 정보, 한글 제거

2.3 데이터 분석 결과

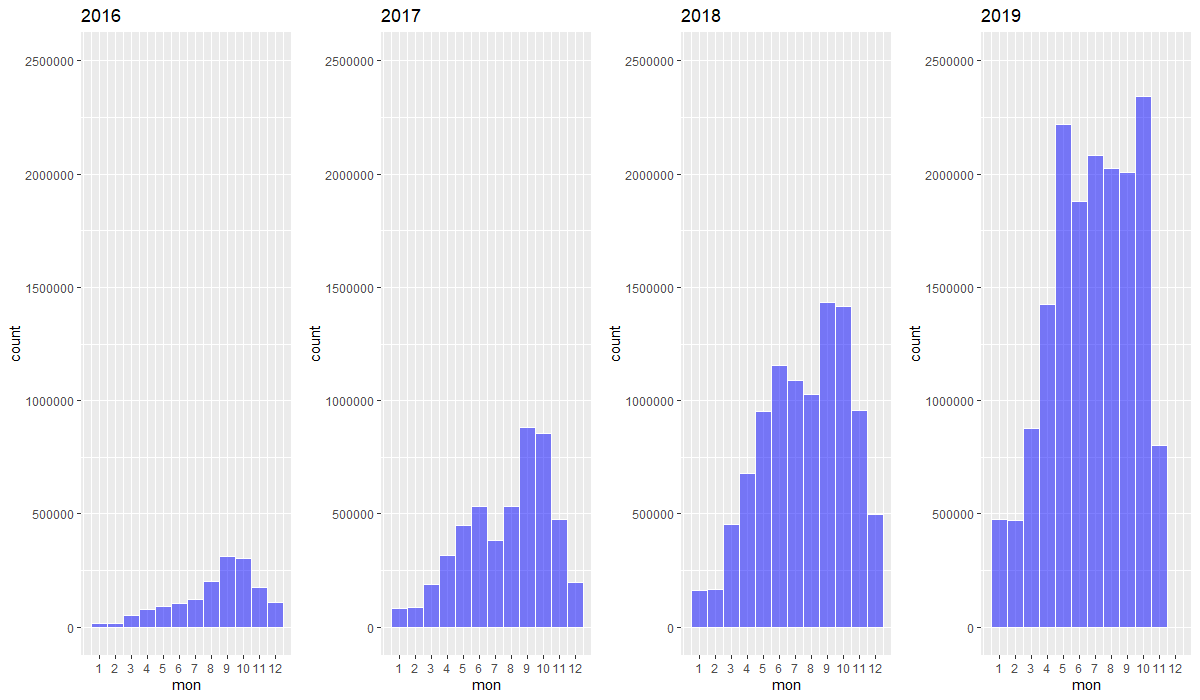


* 2016, 2017, 2018 4개년 비교
  + 연도 별 볼륨 차이가 크므로 년도 별로 나누어서 분석해 볼 필요가 있음

- 따릉이 이용건수는 매년 증가하고 있음을 시각화를 통해 확인할 수 있다.

****

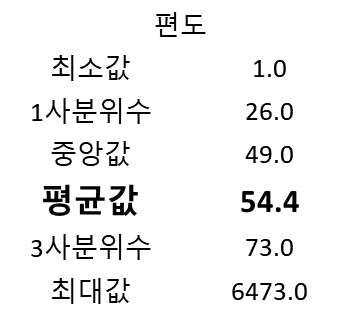
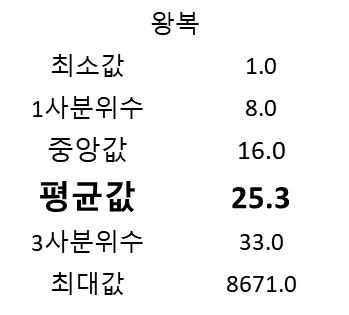
* 2016, 2017, 2018, 2019 4개년 비교
  + 따릉이 이용 시간대는 출근 시간인 8시, 퇴근 시간대인 18-19시에 가장 이용률이 많은 것을 확인 할 수 있다.

****

* 2016, 2017, 2018, 2019 4개년 비교
  + 마찬가지로, 따릉이 이용건수는 매년 대폭적으로 증가하고 있음을 시각화를 통해

확인할 수 있다.

* + 또한 월별로 이용 형태가 다를 수 있음을 예측 할 수 있다.
* **탐색적 분석 시작**



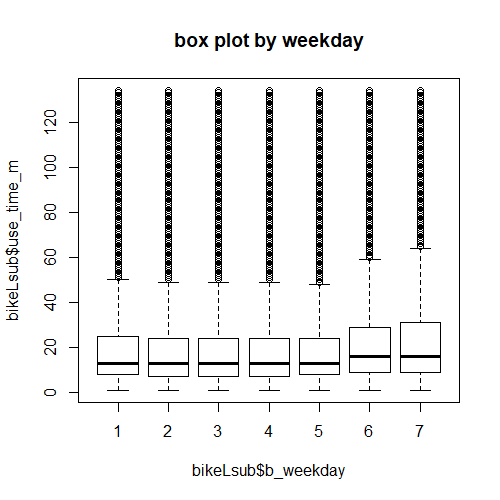
* + 편도 이용자와 왕복 이용자를 구분하여 이용시간 평균을 살펴본 결과, 편도 이용자와 왕복 이용자의 평균 차이가 있을 것으로 예상됨
* **왕복과 편도의 동질성 검정**

-편도와 왕복, 두 집단의 동일집단 여부를 통계적 검증 한 결과 동일집단으로 **볼 수 없다**는 결론





* 주에 따른 왕복집단과 편도 집단의 동질성 검정

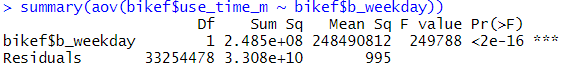


* + ANOVA 분산분석 수행 결과

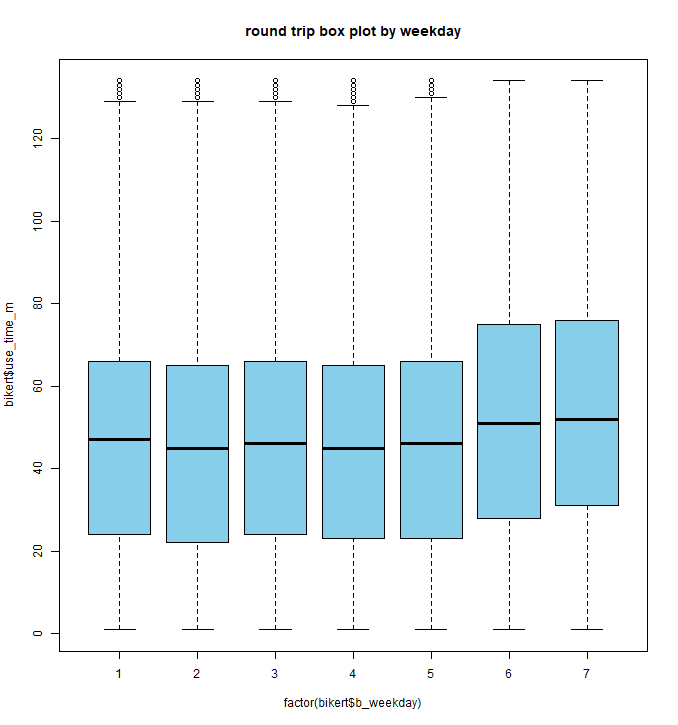
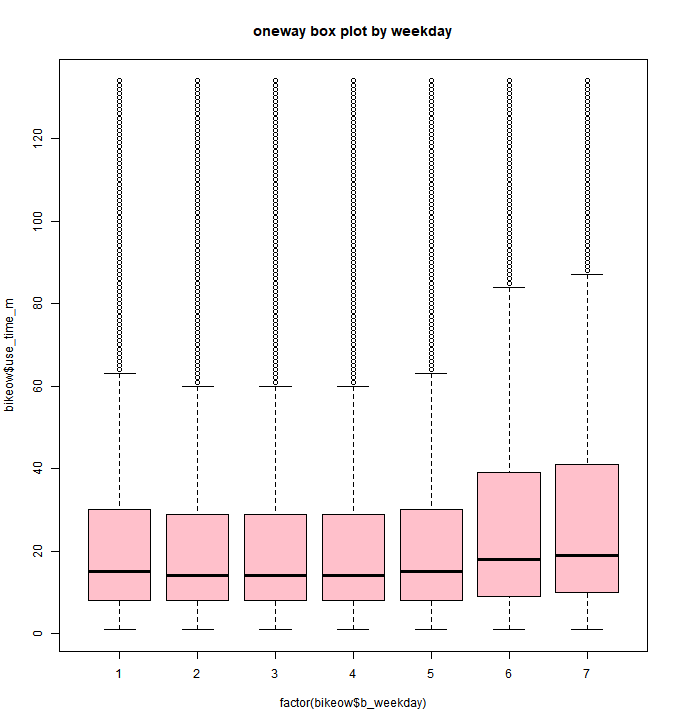
가설: 평일 집단(1~5) & 주말 집단(6,7) 각각의 개체 집단은 서로 연관이 있을 것이다.

결과: P값이 0.05보다 작으므로 귀무가설 기각.

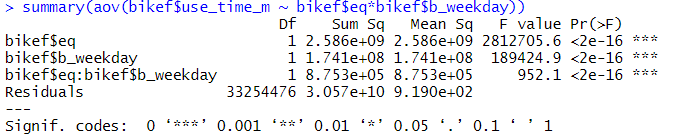
* 즉, 각각 평일, 주말 그룹 내의 값이 서로 연관이 없음을 도출해내었다.



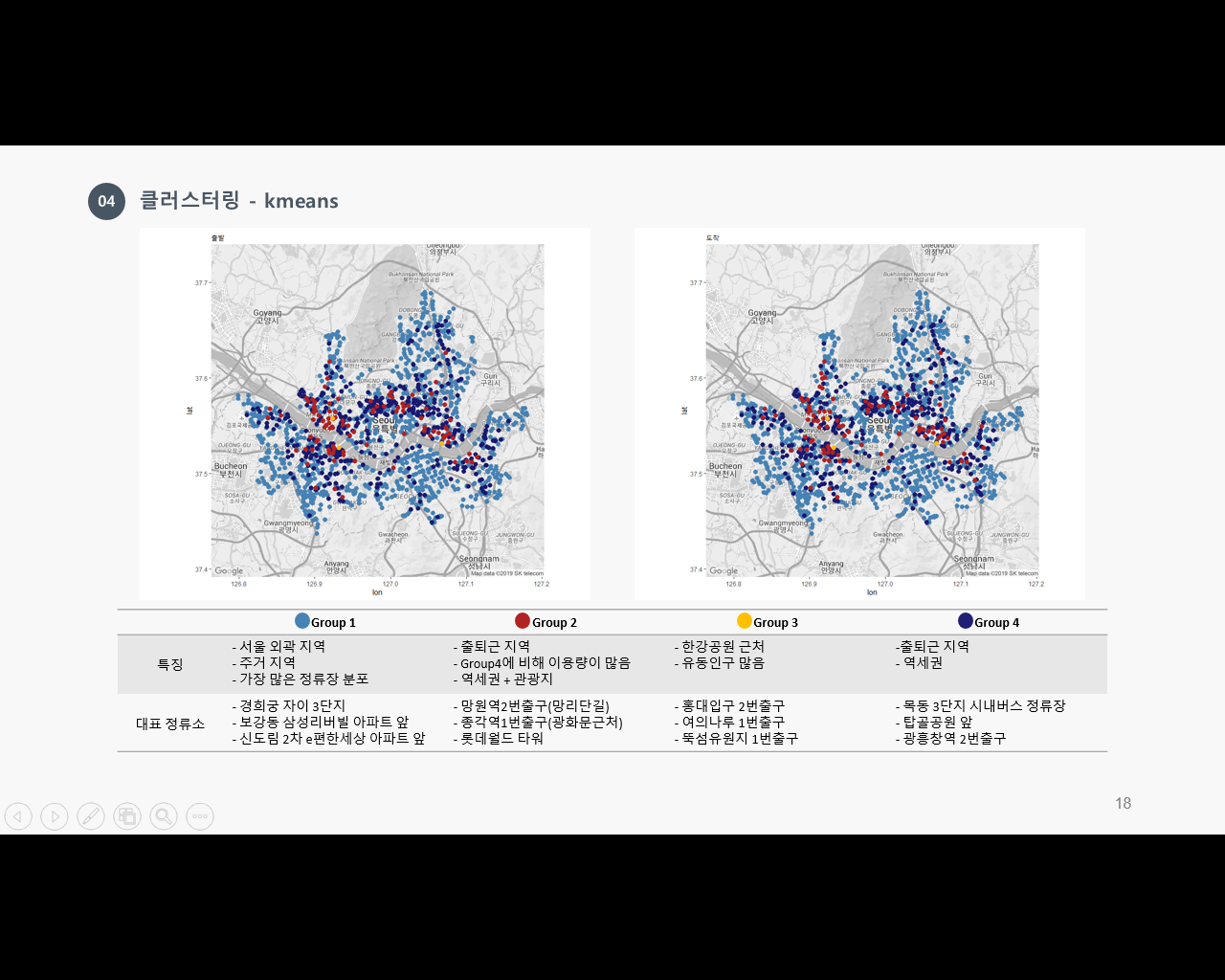
* 분산 분석 결과

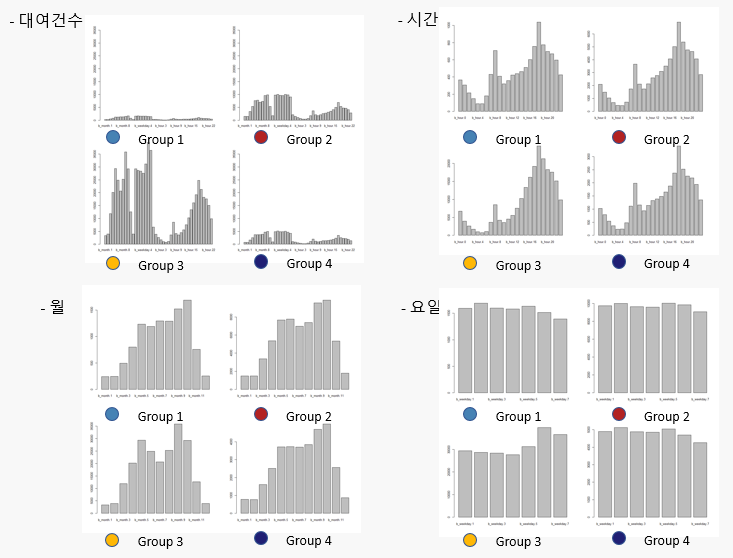


편도와 왕복에 따른 이원분산분석

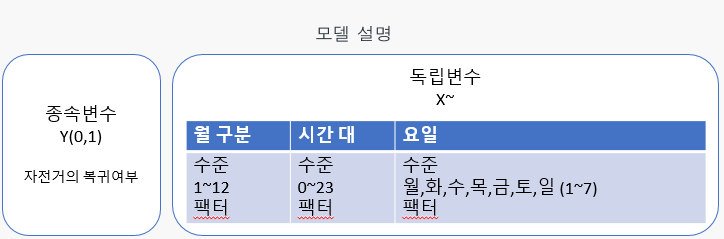


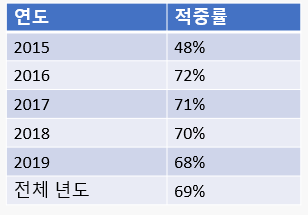
* 이원 분산 분석 결과

* 년, 월, 시간에 따른 왕복 집단과 편도 집단의 동질성 검정도, 주에 따른 동질성 검정과 마찬가지로 P값이 0.05보다 작아 귀무가설 기각.
* 따라서, 평균 이용량에 영향을 미치는 변수들을 도출해 낼 수 있었다.
* **클러스터링 (kmeans)**

****

* + 그룹별로 대여 건수에서 차이를 보이고 있음
  + 3번 그룹의 경우, 계절과 요일에 영향을 많이 받음
  + 3번 그룹은 대여량이 가장 많은 그룹이며, 전체 대여량의 약 2%를 차지함
* **로지스틱 회귀분석**
* 따릉이 이용 패턴에 영향을 줄 것으로 예상되는 요인으로 정류장 위치, 대여 시간대, 월, 요일 등이 있음
* 정류장의 경우 많은 수가 존재하고 정류장 자체를 변수로 사용하기에는 곤란한 점이 존재하므로 주요 정류장을 선정하여 해당 정류장에서 발생하는 사례를 분석
* 연도의 측면에서 보면 대여 건수의 볼륨이 점점 커지는 추세이므로 특정 년도 별로 분리해서 볼 필요가 있음
* 이를 고려하여 대여 건수가 가장 많은 207대여소의 2016년 데이터를 로지스틱 회귀분석에 투입 함





- 207 ' 여의나루역 1번출구 앞‘ 로지스틱 분석 예측 정리표

* 2015년의 266건으로 2016년의 28946, 2017년의 40371, 2018년의 47333, 2019년의 65143 건 에 대하여 로지스틱 분석을 실시해 본 결과
* 207번 정류장에 대해서는 적절한 적중률을 보였으나 다른 정류장에 예측모델을 사용할 경우 거의 예측에 성공하지 못함
* 클러스터별 로지스틱 분석의 경우도 동일 클러스트 내의 정류장을 이용한 예측에서 50% 이상의 예측을 보여주지 못함
* 결과적으로 기존에 주어진 주,월,요일,시간 정보만으로는 예측이 어렵다는 결론

**3. 기대 효과 [따릉이를 통한 혼잡한 교통 활성화]**

3.1 향후 개선 사항

[ 따릉이 편도이용 패턴을 분석하여, 서울시 교통정책 건의 근거 획득 및

따릉이 이용성 개선 목표]

* 모든 대여소를 검토하지 못한 결과로 향후 다양한 대여소를 대상으로 검토해볼 필요가 있음
* 따릉이 자체 데이터 외에 공개적으고 사전적으로 얻을 수 있는 정보를 추가하여 예측력을 향상시켜야 함
* 클러스터 별 특성을 반영한 분석 필요

3.2 기대 효과

* 따릉이 대여시점의 정보만으로 이후에 대여소로 돌아올 것인지 판별

>> 재배치 소요량을 적은 정보량으로 판단

* 대여소 별 특성에 따른 관리로 관리효율을 증대시킬 수 있다.

**4. 개발 후기**

|  |
| --- |
| C:\Users\student\Downloads\IMG_7470.JPG |

|  |  |
| --- | --- |
| 성명 | 후기 |
| 권용진 | 이번 프로젝트 작업을 통해 몇천만건의 데이터를 실제로 다뤄볼수 있었습니다. 몇천만건의 데이터를 R로 작업할 때, 데이터량이 많아서 기다림의 시간도 필요했었습니다. 원하는 시각화를 얻기 위해 진행한 구글링 작업에서는 필요한 정보를 찾는 것도 어려웠지만, 가지고 있는 데이터를 얻은 정보에 적용하는 것도 쉽지 않았습니다. 하지만, 데이터 정제과정을 거쳐 시행착오를 거듭한 끝에 원하는 시각화 결과를 냈을 때 보람찼습니다 추가적으로 더 도출해 낼 수 있는 시각화 작업을 해 볼 예정입니다. |
| 신현정 | 항상 버스를 타고 대로변을 지나갈 때, 곳곳에 따릉이가 구비되어 있는 것을 많이 봤었습니다. 실제로 회원가입을 통해 한번 밖에 따릉이를 이용해보진 못했지만 운동 뿐만 아니라 교통수단으로 매우 편리하다는 생각을 했었습니다. 아침과 저녁, 멀티캠퍼스에 오기 위해 혼잡한 지하철을 이용 할 때, 따릉이의 환승 서비스가 도입되어 인구 밀집의 분산을 통해 교통체증이 조금 해소 되기를 바랬습니다. 이번 프로젝트를 통해 직접 수십만개의 데이터 전처리 과정을 통해 기다림과 배움을 얻었고 최종 프로젝트까지 더욱 더 좋은 기반이 될 수 있었던 프로젝트이었던 것 같습니다. |
| 유지수 | 이번 프로젝트를 통해 처음으로 몇천만건의 대량 데이터를 가공하는 경험을 할 수 있었습니다. 정제되지 않은 대량의 데이터를 전처리 하면서 예상치 못했던 문제점들이 발견되었고 그러한 문제를 하나하나 해결하는 과정에서 많은 배움을 얻을 수 있었습니다. 여러 분석과 시각화를 진행하면서, 시각화의 종류와 그에 따른 데이터형태는 무엇인지에 대해서 알 수 있게 되었습니다. 이론적으로만 알고 있던 검정방법들도 직접 적용해보면서 어떤 때에 검정이 필요할지를 느낄 수 있었습니다. 앞으로 프로젝트를 진행하게 될 때, 이번 프로젝트에서 얻은 배움을 통해 더 폭넓은 시야를 가지고 문제에 대해 고민할 수 있을 것 같습니다. |
| 정재원 | 학습할 때와는 다르게 정제되지 않은 대용량의 데이터를 다루다 보니 문제가 많이 생기는 것을 확인할 수 있었습니다. 이 과정에서 데이터의 처리속도를 높이는 요령이 필요함을 느끼고 여러 가지로 궁리해 보았습니다. 특히 처리량이 많아지면서 컴퓨터가 다운되는 경우가 자주 발생하여 이 부분에 많이 고심하게 되었습니다. 프로젝트 진행 과정에서 저음 의도한 데로 잘 풀리지 않는 점도 많았으나 팀원들과 상의하며 다시 방향을 잡아갈 수 있었습니다. 협력을 하면서 혼자서는 찾아낼 수 없던 오류를 서로 잡아 줄 수 있었던 점도 좋았습니다. |