<table>
<tr><td colspan="2">팀 명</td><td colspan="3">극데노(극한 데이터 노예)</td></tr>
<tr><td rowspan="4">대표자</td><td>성 명</td><td>유제우</td><td>생년월일</td><td>1999.05.31</td></tr>
<tr><td>주 소</td><td colspan="3">주민등록상 주소 기재</td></tr>
<tr><td>전화번호</td><td>01087387616</td><td>이 메 일</td><td>wpdn115@naver.com</td></tr>
<tr><td>소 속</td><td colspan="3">강남대학교(데이터사이언스 전공)</td></tr>
<tr><td rowspan="6">팀원 정보</td><td colspan="2">성명</td><td colspan="2">소속</td></tr>
<tr><td colspan="2">이서영</td><td colspan="2">강남대학교(데이터사이언스 전공)</td></tr>
<tr><td colspan="2">이현중</td><td colspan="2">강남대학교(데이터사이언스 전공)</td></tr>
<tr><td>분석 주제</td><td colspan="3">서울시 시민의 지하철, 버스의 대중교통 데이터, 30분 단위 이용 통계, 여러 서울시 공간데이터를 활용해 사람들의 출발지와 도착지 정보를 이용하여 시간별 혼잡 상황을 파악하고, 혼잡도를 예측한다. 그 후, 인원이 몰리는 특정 역의 이유를 찾고 사람들의 출발역, 도착역 정보를 파악하여 균등한 객차 혼잡비율을 조절해 시각화를 통한 인사이트를 도출한다.</td></tr>
<tr><td>멘토링 분야</td><td colspan="3">■ 인구/가구 □ 보건/복지 □ 문화/관광 ■ 교통/물류 □ 환경/기상
□ 교육 ■ 도시/지역 □ 산업/고용 □ 경제/금융 □ 행정/안보
□ 재난/안전 □ 교육 □ 과학기술/에너지</td></tr>
<tr><td rowspan="2">활동 계획</td><td>분석 내용</td><td colspan="3">머리글
지하철은 버스와 택시에 비해 많은 승객들을 안전하고 신속하게 수송할 수 있는 미래 지향적인 교통수단이다. 지하철 이용자의 증가에 따른 혼잡도 증가는 지하철을 쾌적하게 이용할 수 있는 시민들의 권리를 저해하는 요인 중의 하나이다. 따라서 지하철 내의 혼잡도 예측은 승객의 이용 편의성과 쾌적성을 극대화할 수 있는 방법 중 하나이다. 본 분석에서는 기존의 지하철 혼잡도를 python 을 통해 예측하고 matplotlib 과 seaborn 을 통한 혼잡도를 시각화한다.

주제어
혼잡도 분석 예측, 빅데이터, KDD분석 방법론, 공간데이터

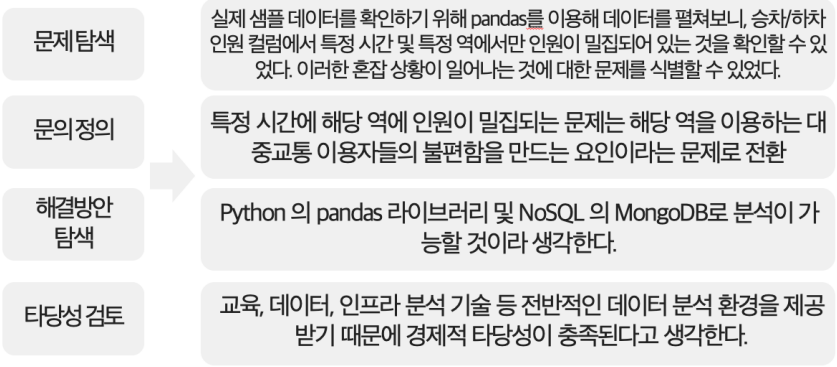
[서론]
지하철은 버스와 택시와 같은 교통수단에 비해 많은 승객들을 목적지까지 안전하고 신속하며 정확하게 원하는 지점으로, 대량 수송할 수 있어 미래 지향적인 교통수단이라 할 수 있다. 이러한 지하철 운행 특성에 따라 출퇴근 시간뿐만 아니라 일상적인 대중교통으로서의 수요 또한 급증하고 있다. 1995년 134만 명을 기준으로 2010년236만 명으로, 해가 거듭할 수 록 지하철 이용인원이 꾸준히 증가하고 있고, 6대 도시권에서 1시간 이상 출퇴근하는 이동 인구가76% 증가하였다[1].
승객 혼잡도를 해결하고 승객의 안전과 편의성을 제공하기 위해 CCTV를 이용하여 혼잡도를 측정하고, 공공 데이터 분석을 통해 새로운 대안을 제시하는 등의 노력도 있지만[2], 이보다 앞서 개인정보나 초상권등과 같은 인권이 보호되는 범위 내에서 보장되도록 해야한다.

서울시와 서울교통공사는 지하철 혼잡 정보를 알려 승객 분산을 유도하는 방안인 '혼잡도 사전예보제'도 시행하기로 했다. 이는</td></tr>
</table>

방송·SNS 등 각종 매체, 지하철 역사 및 열차 안내방송, 전광판 등을 통해 지하철의 시간대별·호선별 혼잡 정보를 제공해 시민의 자율적인 이용 분산이 이뤄지도록 하는 것이다[3].
추가로 혼잡도 사전예보를 위해 여러곳에서 제공되는 지하철 관련 데이터를 활용하여, '또타지하철'과 같은 어플이 개발되고 있는 시점이다.
혼잡도 분석이 중요한 이유는, 효율적인 차량 운영 계획 때문이다. 각 운영기관에서 혼잡도의 활용은 첫째, 환승현황 및 혼잡도 결과를 토대로 제반도시 철도 시설물의 안전관련 사항과 역운영의 참고자료, 둘째, 도시철도를 이용하는 승객에 대한 안전 확보와 서비스 향상 자료, 그리고 세번째는 교통량의 변동량 추이를 분석하여 승객 서비스 수준을 유지하기 위한 자료 등이다.

[**본론**]

출퇴근 시간에 사람들이 붐벼서 못타는 경우가 많거나 지하철 배차 간격이 합당하지 않아 여러 불편함을 호소한 경우가 많았기 때문에, 빅데이터캠퍼스의 데이터 현황에서 추천데이터의 항목중'대중교통 30분단위 이용 통계'데이터가 이목을 끌었다. 본 분석은 해당 관련 데이터 세트를 가지고 '하향식 접근법'을 통해 문제를 탐색하고 해당 문제를 정의 및 해결방안 탐색을 진행하기로 한다.



'대중교통 30분 단위 이용 통계'데이터를 살펴보니 정형데이터인 것을 확인할 수 있었다.

1. 데이터 준비

가. 데이터 선택

1) 서울시 대중교통 대한 샘플 데이터는 총 6개 였고, 최근 대부분의 대중교통 이용자들이 교통카드를 사용하는 경우가 비일비재하기 때문에 1회권에 데이터 세트는 배제할 생각이다.(sample 데이터를 보고 판단한 것이기에 추후 빅데이터 확인 후 변경가능.) 따라서 그 중 우리가 이용할 데이터는'서울시 지하철 30분 단위 이용 통계'이다. 이 데이터는 각 컬럼별로 날짜 및 시간, 호선명, 역명 그리고 승차/하차에 대한 인원에 대한 정보가 들어 있다.

2) 추가로 해당 역에서 왜 인원이 많이 내리는지 인과관계를 도출하기 위해 '서울시 대중교통시설 위치정보'의 데이터를 이용할 것이다. 이 데이터는 NoSQL의 MongoDB툴을 이용할 것이다.

가. 데이터 전처리

1) 데이터 분석에 앞서 데이터에 포함되어있는 Noise, Outlier, Null을 파악하고 제거해야한다. 예를 들어 '서울시 지하철 30분 단위 이용통계'의 데이터 세트에서 '역ID'컬럼은 추후 데이터 분석 및 예측에 활용되지 않을 컬럼이기 때문에 해당 열을 삭제시킬 것이고, 결측치가 존재한다면 특정 값으로 대체를 시킬 것이다.

2) 승차인원/하차인원 컬럼에는 매우 다양한 범위의 숫자가 기재되어 있는 것을 확인하였고, 적은 숫자의 범위를 가진 인덱스는 데이터 분석에 있어서 활용되지 않을 것 같아 승/하차 인원이 100명 이하인 인덱스는 모두 삭제를 진행하였다.

2. 데이터 분석

나. EDA

1) 데이터 전처리를 통해 정제된 데이터를 보고 이를 다양한 각도에서 관찰하고 이해를 우선으로 할 것이다. 한마디로 데이터를 분석하기 전에 그래프나 통계적인 방법으로 자료를 직관적으로 바라 볼 것이다. 이 과정에서 시각화 또한 사용할 것이다. 이를 통해서 패턴을 발견하거나 데이터의 특이성을 확인하거나 통계와 그래픽 (혹은 시각적 표현)을 통해서 가설을 검정할 것이다.

다. 데이터 마이닝

1) 해당 역에 회사나 주요시설(맛집, 영화관, 학교, 병원 등)이 많은지를, 역에 내리는 사람의 수를 통해 인과관계를 도출한다. 호선별 어느 시간대에 이용 고객이 많은지 대중교통 데이터를 통해 사람들의 출발지와 도착지 정보 이용 하여 시간별 혼잡 상황 파악 한다. 혼잡도를 미리 예측 하고 몰리는 특정 역 위치 이유 찾고 사람들의 출발역, 도착역 정보를 파악하여 균등한 객차 혼잡비율 조절 한다. 이런 방식으로 기존에 알려지지 않은 패턴이나 규칙을 찾아낼 것이다.

라. 시각화

1) 빅데이터는 대규모 데이터가 존재하기 때문에 많은 양의 데이터를 직관적으로 전달할 수있는 방법이 필요하다. 따라서 데이터 분석을 통해 얻는 정보들을 python의 matplotlib, seaborn 라이브러리를 이용하여 시각화를 진행할 것이다.

2) 데이터의 분포를 확인하기 위한 파이차트, 트리맵을 활용할 계획이다. 데이터의 컬럼별로 상관관계를 확인하기 위해 히트맵을 이용할 것이다.

3. 예측

끝으로 EDA 밑 데이터 전처리가 완료되어 데이터가 어느정도 정제 되었다면, Multiple Regresstion Analysis(다중회귀), Tree Ensemble(트리의 앙상블)등 다양한 회귀기법을 통해, 시간에 따른 지하철 혼잡도를 예측할 것이다. 이를 통해 지하철 승객들의 이용 편의성과 쾌적성을 극대화하는 방안을 마련해 보고 싶다.

[결론]

빅데이터를 통해 사람들의 출발지와 도착지 정보를 이용하면 실시

		간 혼잡 상황을 알 수 있으니, 각 구간 혼잡도를 미리 예측해서 예측된 혼잡도를 바탕으로 효율적인 추천을 유도하고, 전체적인 혼잡도를 균등하게 배포가능하다. 또한 지하철 객차 내 혼잡도는 사람들의 동선 최소화하려는 심리적 요인으로 인해 환승통로, 계단, 엘레베이터 같은 특정 지점이 다른 지점에 비해 혼잡도가 상당히 높다. 이러한 심리적 요인과 사람들의 출발역, 도착역, 환승역 정보를 파악하여 균등한 객차 혼잡비율을 조절한다. 지하철 간의 환승은 제대로 파악되지 않아 해당 역의 승하차 인원만으로 혼잡도를 정확히 예측하기 어렵고 이를 해결해 보고 싶다. 각 지역 지하철에 대한 다양한 정보 데이터베이스를 생성하고, 각 역별 혼잡도 예측을 위한 다중 회귀모형 구축하고 싶다. ---------- REFERRENCES [1] Yong-Hyun Cho, “Metropolitan commuting time in half”, Koera Railroad Research Institute, 2013. [2] Keun-Won Kim, Dong-Woo Kim, Kyoo-Sung Noh, Joo-Yeoun Lee, “An Exploratory Study on Improvement Method of the Subway Congestion Based Big Data Convergence”, JOURNAL OF DIGITAL CONVERGENCE, Vol.13, No.2, pp.35-42, 2015 [3]https://terms.naver.com/entry.naver?docId=5945566&cid=43667&categoryId=43667
	활용 예정 데이터	1. 서울시 지하철 30분단위 출발-도착 데이터 2. 서울시 지하철 30분단위 이용 통계 3. 서울시 우수중소기업 공간데이터 4. 서울시 주요시설과 집객시설 공간데이터 5. 서울시 지하철 시간대별 승객 수 6. 서울시 상주인구 공간데이터
	사용할 데이터 분석 및 시각화 방법	[데이터 분석] - python pandas, numpy등 라이브러리를 이용한 데이터 전처리 및 분석 - NoSQL MongoDB -> index의 2dsphere이용 [시각화] - python matplotlib, seaborn 라이브러리를 이용한 데이터 시각화
	기 타	저희가 작성한 분석 주제 및 내용을 바탕으로 머신러닝 알고리즘을 이용해 시간대별 지하철 이용객의 승하차 인원을 예측해보고 싶습니다. 또한 한정적인 데이터를 가지고 다양한 인사이트를 도출 할 수 있는 다양한 방법을 익히고 싶습니다. 전문가님의 커뮤니케이션 능력을 배우고 싶습니다.

2022 서울특별시 빅데이터캠퍼스 멘토링

참 가 신 청 서

※ 활동계획은 멘토링 진행 시 변경가능

2022. 05. 14.

대표자: 유제우 (인)

서울특별시장(빅데이터캠퍼스) 귀하

개인정보 수집 · 활용 동의서

• **【개인정보의 수집·이용 목적】**
: 2022년 서울특별시 빅데이터캠퍼스 멘토링 참여 및 투입인력들의 기초자료를 확보하여 심사과정 및 선정 이후 원활한 사업수행관리

• **【개인정보의 수집항목·이용 목적】**

수집·활용 항목	수집·활용 목적	이용기간 및 보유기간
성명, 이메일, 생년월일 연락처, 주소	빅데이터캠퍼스 멘토링 운영 등	신청일로부터 5년간 (2027 12월까지 활용·보관)

- 수집한 개인정보는 보유 기간 이후 폐기됩니다.
- 개인정보는 멘토링 참가 및 활동 이외에 이용하거나 제3자에게 제공하지 않습니다.
- 참가자는 개인정보보호법 제15조 제2항 제4호에 따라 개인정보 수집 및 이용에 관하여 거부할 수 있습니다. 그러나 동의를 거부할 경우 원활한 서비스 제공에 일부 제한을 받을 수 있습니다.

위 내용은 모두 사실이며, 허위 기재일 경우 참가에서
자동으로 취소됨을 사전에 미리 양지하고 있습니다.

2022년 05월 14일

신 청 인 : 유제우()

서울특별시장 귀 하