|  |  |
| --- | --- |
| **프로젝트 결과 보고서** | |
| 작 품 명 | Super Crazy |
| 팀 명 | Billy |
| 제 작 기 간 | 2018년 11월 05일 ~ 2018년 11월 09일 |
| 프로젝트 매니저 | 송재환 (인) |

|  |  |  |  |
| --- | --- | --- | --- |
| 조 원 | | | |
| No. | 성 명 | 연 락 처 | E-mail |
| 1 | 김민주 | 010-8594-3921 | mjk173a@gmail.com |
| 2 | 김태리 | 010-3382-6758 | serendipity.in.you@gmail.com |
| 3 | 박정훈 | 010-7196-8830 | wndwo600@naver.com |
| 4 | 송재환 | 010-5699-7956 | holibaby@naver.com |

1. **Team 소개**

* **역할**

김민주 : 크롤링 / 시스템 환경구성 / 보고서 작성

김태리 : 데이터 전처리 및 데이터 분석

박정훈 : 데이터 전처리/후처리 및 데이터 분석 / 시각화

송재환 : 프로젝트 일정 및 과정 / 결과물 관리 / 데이터 전처리/후처리 및 데이터 분석 / 발표 자료 제작

1. **Project 주제 : 고객데이터 분석을 통한 신규 공연 브랜드 기획**

주말을 의미 있게 보내고 싶어 하거나 여가문화생활에 공연을 찾는 고객의 수가 늘고 있습니다.

여가 생활이 전 연령대에 걸쳐서 화두인 가운데 여러가지 문제 연구소장 김정운교수가 쓴 <잘 노는 사람이 성공한다.> 는 책이 꽤 오랫동안 베스트셀러에 있는 것을 보아도 점점 문화생활에 대한 수요가 커지는 것을 알 수 있습니다.

이번 프로젝트를 통해 기존 고객들의 보다 나은 문화생활을 위한 공연 브랜딩을 기획해보았습니다.

기존의 고객들이 이런 부가적인 서비스를 통해 만족을 얻어 브랜드 충성도를 높일 수 있으며 더 나아가 신규 고객을 유치 하는데에도 기여할 수 있을 것으로 보입니다.

1. **Project 개요**

* **소개**

수집한 데이터 ( 공연 관객의 리뷰 공감 수, 부동산 거래 가격, 공연 티켓 가격, 네이버 검색어 트렌드) 를 기반으로 공연 기획과 마케팅분야에 활용하기 위해 이와 같은 서비스를 기획했습니다.

* **주제 선정 이유**

수요가 늘어나고 있는 문화생활시장에 맞춤형 서비스와 더불어 고객 맞춤 문화컨텐츠를 통해 브랜드의 인지도와 시장의 영향을 확보하기 위해 선정했습니다.

1. **활용 범위**

* **공연 기획**

단조로운 일상을 벗어나 주말만은 모두가 특별하게 보내고 싶어합니다. 뿐만 아니라 의미 있게 보낸 시간으로 다음주를 살아갈 에너지를 얻기도 합니다. 매년 다양한 형태의 공연을 찾는 고객의 수가 늘어나고 있는 가운데 고객 맞춤 공연기획에 있어 예산과 타겟층을 설정하고 수익률과 예산 두가지 토끼를 잡을 수 있게 데이터를 활용할 예정입니다.

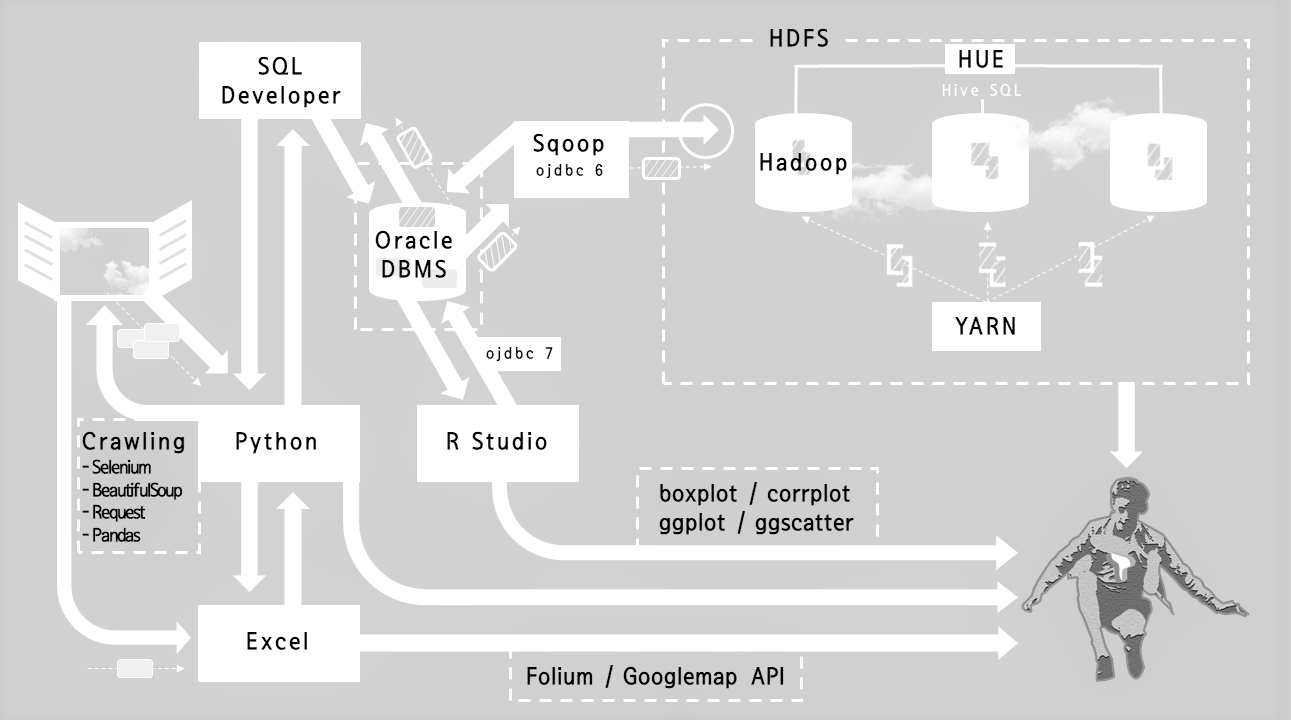
* **마케팅분야**

사용자들이 이 서비스를 이용하면서 축적된 데이터들을 기반으로 각 고객의 취향을 파악하여 계절과 개인 이벤트와 같은 고객 맞춤형 공연 추천이 가능해지고, 공연 기획사에서는 축적된 데이터를 기반으로 공연 선택에 있어 사용자별 선호도 측정이 가능해져서 고객별로 맞춤 공연 추천 서비스가 가능합니다.

1. **개발 내용**

* **시스템구성도**

웹에서 크롤링, 공공데이터 수집을 통해 데이터를 가져와 엑셀, 파이썬, R을 이용해 데이터 전처리 후 적재하고, 같은 툴을 이용해서 분석하고 시각화 하는 과정



* 크롤링 : Selenium과 chrome driver를 이용해 브라우저를 조정, Request 라이브러리로 html 문서를 받아오고 Beautiful Soup를 사용해 가져온 html 문서를 정형화하고 특정한 태그 부분을 가져온 후 파이썬 문자열 처리 내장함수를 이용해 간단한 전처리. 전처리 된 리스트를 Pandas로 데이터 프레임으로 만든 후 csv 파일로 출력.
* Oracle DBMS : 리눅스 기반 가상서버 (192.168.119.119:1521)에 위치하고 윈도우 환경의 SQL Developer, R studio 와 연결하여 데이터 전처리, 분석, 시각화 진행.
* HDFS : 고객 데이터를 적재하고 Hue, Hive를 통해 HDFS 시스템 내부에서 고객 데이터 전처리 후 처리된 데이터를 Oracle DBMS 옮겨 사용, Oracle db와의 연결은 Sqoop을 이용. 데이터를 받아올 때 Yarn을 통해 Map reduce 과정을 거치고 각 가상서버에 분산저장됨.
* 시각화 : R studio 중심으로 고객정보데이터 분석, 시각화 하였고 웹을 통해 수집한 서브데이터들은 excel과 python을 통해 하였음
* **사용기술**

|  |  |  |  |
| --- | --- | --- | --- |
| **구분** | **이름** | **버전** | **목적** |
| **사용 툴** | **SQL Developer** |  | **DBMS의 데이터 전처리, 분석** |
| **RStudio** |  | **DBMS의 데이터 전처리, 분석, 시각화** |
| **Jupyter Notebook** |  | **크롤링, 데이터 전처리** |
| **Excel 2016** |  | **데이터 전처리, 시각화** |
| **jdk 1.8** | **1.8.0\_181** |  |
| **jdk 1.7** | **1.7.0\_67** |  |
| **API** | **OJDBC 7** | **7** | **RStudio, Oracle DB 연결** |
| **OJDBC 6** | **6** | **Hadoop, Oracle DB 연결** |
| **DBMS 및 기타 시스템** | **Oracle DBMS** | **4.3.5** | **고객데이터, 수집데이터 저장** |
| **CDH** | **5.15.1** | **빅데이터 관리 시스템** |
| **HDFS** | **2.6.0-cdh5.15.1** | **Hadoop 빅데이터 분산 적재** |
| **Yarn** | **2.6.0-cdh5.15.1** | **Map reduce, 시스템 자원 관리** |
| **Hive** | **1.1.0-cdh5.15.1** | **빅데이터 관리 시스템에서 sql 사용** |
| **Hue** | **4.2** | **HDFS 사용자 편의 관리** |
| **Sqoop 1 Client** | **1.4.6-cdh5.15.1** | **RDBMS 와 Hadoop 기반 적재시스템간의 연결** |
| **라이브러리** | **Selenium / webdriver** |  | **크롤링** |
| **bs4 / BeautifulSoup** |  |
| **requests** |  |
| **time** |  |
| **pandas** |  |
| **googlemaps** |  | **지도생성** |
| **import googlemap\_key** |  |
| **folium** |  |
| **nltk** |  | **자연어 처리** |
| **ggplot2** |  | **그래프 그리기** |
| **plotly** |  | **그래프 그리기** |

* **데이터**

- 인터파크 리뷰, 인터파크 뮤지컬 관람평 크롤링 데이터

: raw\_data 150행, 자연어처리 후 16926행

- 아파트(매매)\_실거래가 / 국토교통부 실거래가 공개시스템

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 1월 | 2월 | 3월 | 4월 | 5월 | 6월 | 7월 | 8월 | 9월 | 10월 |
| 12566 | 9169 | 9403 | 4480 | 4702 | 5237 | 7041 | 14980 | 6610 | 1126 |

- 공연 현황 / 네이버 공연 크롤링

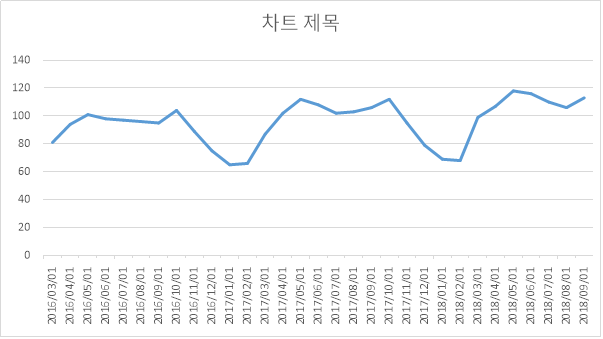
: 클래식 160행, 콘서트 150행, 뮤지컬 70행, 연극 166행

- 카드승인내역 / 여신금융협회

: 카드 승인실적 보고서 58개

* **분석**

- 카드승인내역 그래프



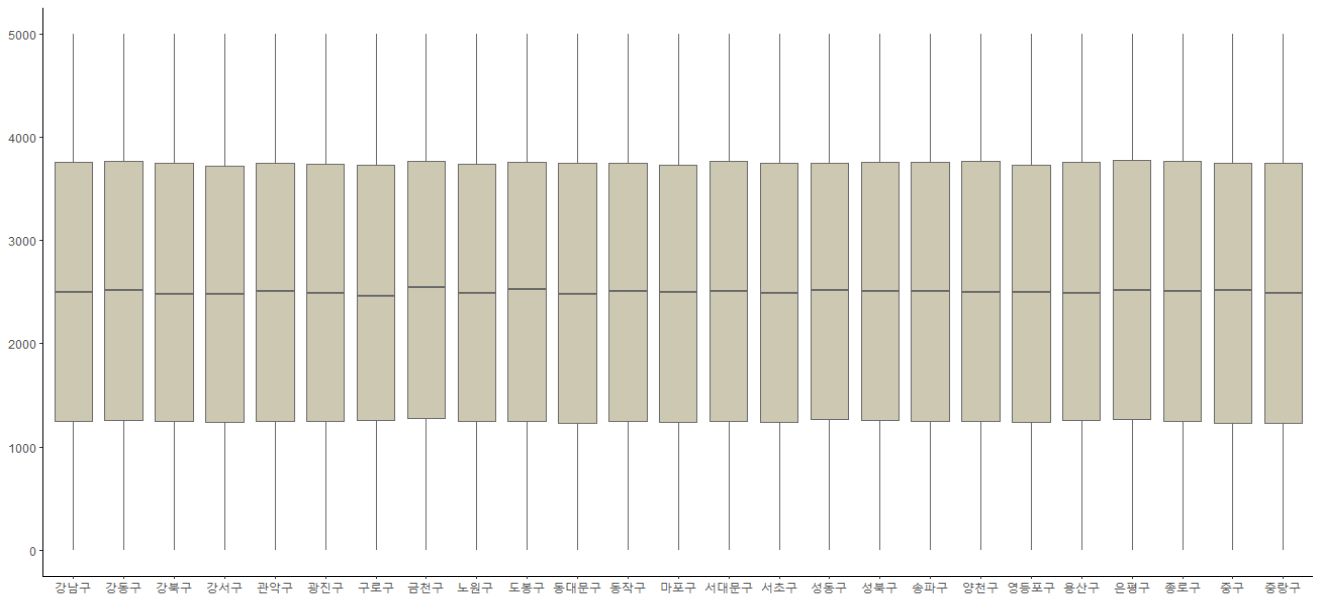
**(1) 고객데이터 분석**

- 서울 구역별 신용 한도를 boxplot함수로 구현

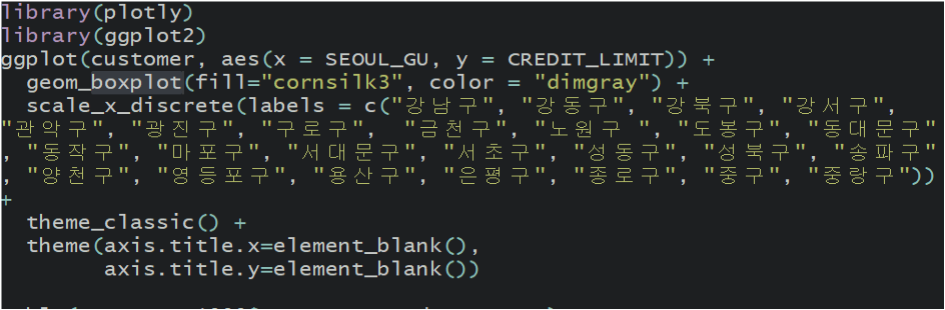
데이터 종류 : 전처리 되어 있는 customer 테이블

데이터 개수 : 887180row(서울 거주 인구 수), 2col (CREDIT\_LIMIT, SEOUL\_GU)

신용한도를 서울 구역별로 나눠서 표현, x값은 서울 구, y값은 신용한도, ggplot2패키지에 있는 boxplot함수를 이용해서 시각화 시킴

****

\* 코드

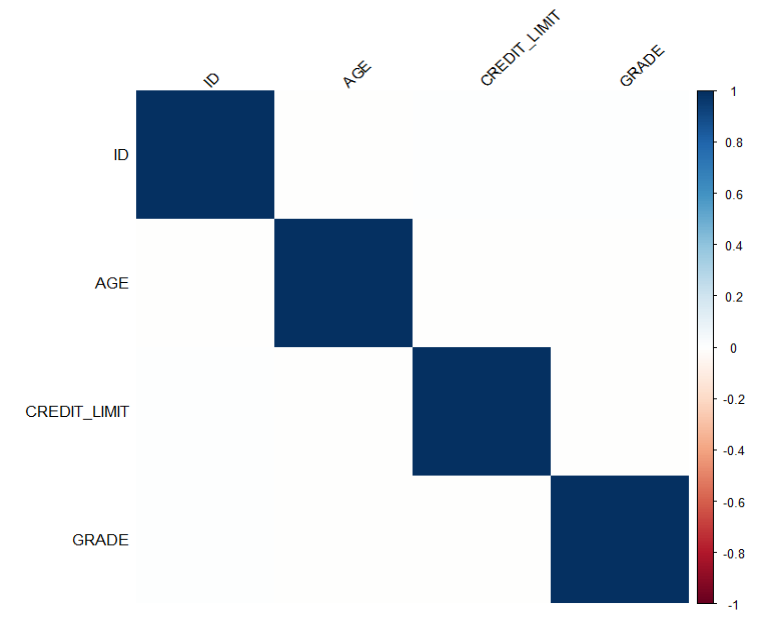
****

- 수치형 데이터의 상관관계 분석(-1≤r≤1) (corrplot함수로 구현)

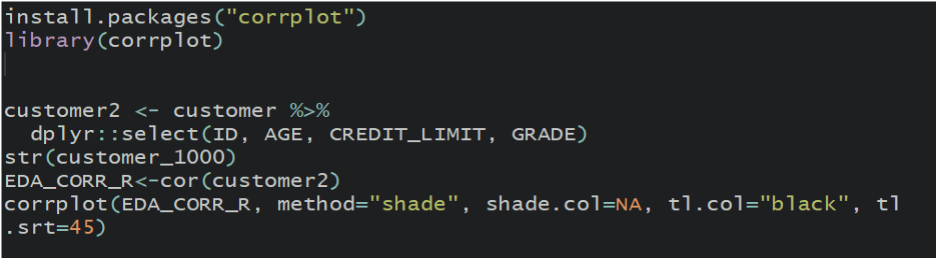
데이터 종류 : 전처리 되어 있는 customer 테이블

데이터 개수 : 887180row(서울 거주 인구 수), 4col (ID, AGE, CREDIT\_LIMIT, GRADE)

customer테이블의 수치화 된 데이터를 비교하기 위해서 corrplot패키지의 corrplot함수를 사용해서 시각화 했다. x와 y값을 4개의 변수로 지정하고 행렬모양으로 구현시켜서 서로 다른 변수를 비교할 수 있게 했다. 결과값은 같은 변수를 제외한 모든 값이 0이어서 4개의 변수는 서로 관련이 없는 것으로 결론을 내렸다. GRADE column은 지역별 부동산 매매가를 이용해서 1~25등급으로 나눴다.



\* 코드

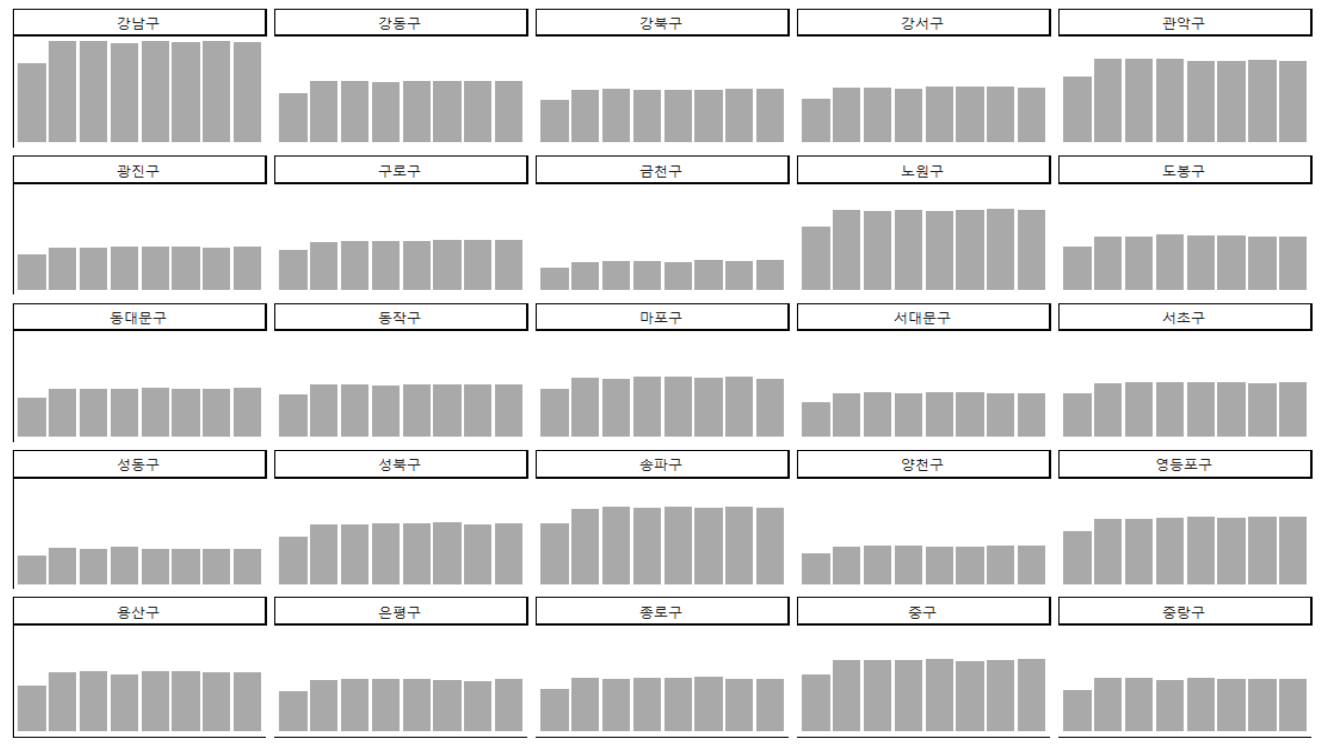


- 서울 구역별 인구 분포(연령을 8단계로 나눔)(ggplot중 geom\_bar로 구현)

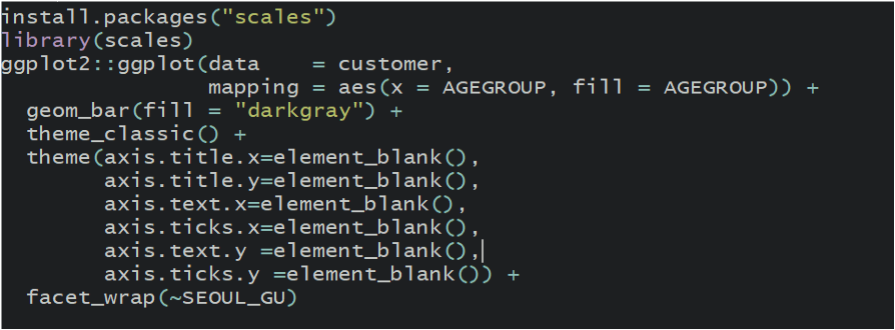
데이터 종류 : 전처리 되어 있는 customer 테이블

데이터 개수 : 887180row(서울 거주 인구 수), 2col(AGEGROUP, SEOUL\_GU)

서울 지역별 연령분포를 파악하고자 ggplot패키지에 들어있는 geom\_bar함수를 사용해서 시각화 시켰다. x값은 AGEGROUP이고, y값은 AGEGROUP 빈도 수이다. AGEGROUP column은 AGE 값을 8등급으로 나눠서 전처리 했다. 결과 값은 제일 어린 성별을 제외하고 나머지 연령이 고르게 분포되어 있다는 것이다.



\* 코드



- 구별거주인구별 구별부동산매매가(ggscatter함수로 구현)

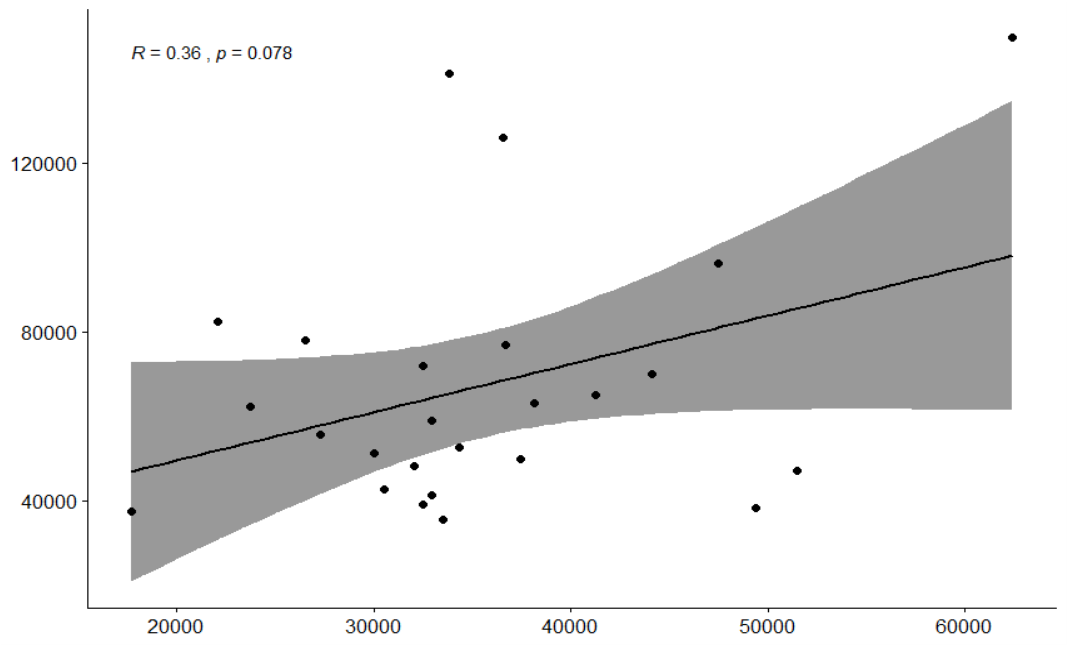
데이터 종류 : 월별 문화 공연 관람에 쓰인 카드 사용 내역

데이터 개수 : 31row(2016년 3월부터 2018년 9월 까지 월 단위 개수) \* 2column(날짜(월 단위), 카드 사용 금액(단위 백 억원))

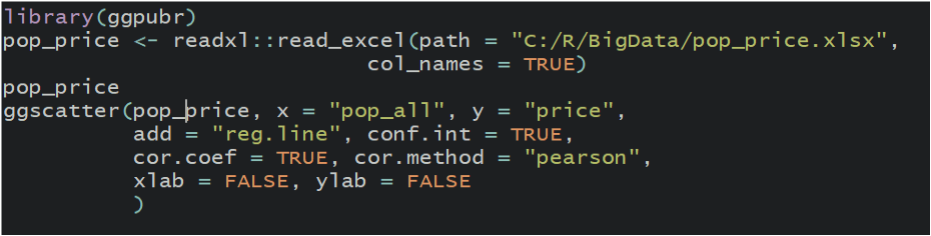
데이터 종류 : 구별거주인구별 구별부동산매매가

데이터 개수 : 25row(서울 구역 개수), 2col((부동산 매매가, 거주 인구수))

구별거주인구와 구별부동산매매가의 관계를 ggpubr 패키지에 들어있는 ggscatter함수를 사용해서 시각화 시켰다. x 값은 구별 거주 인구수이고, y 값은 구별 부동산 매매가이다. R값이 0.36으로 나온 것으로 볼 때, 두 변수가 어느 정도 관계가 있는 것으로 설명할 수 있다. 다시 말해, 구별거주인구가 많을 수록 부동산 매매가가 어느 정도 높을 수 있다는 것으로 설명할 수 있다.



\* 코드

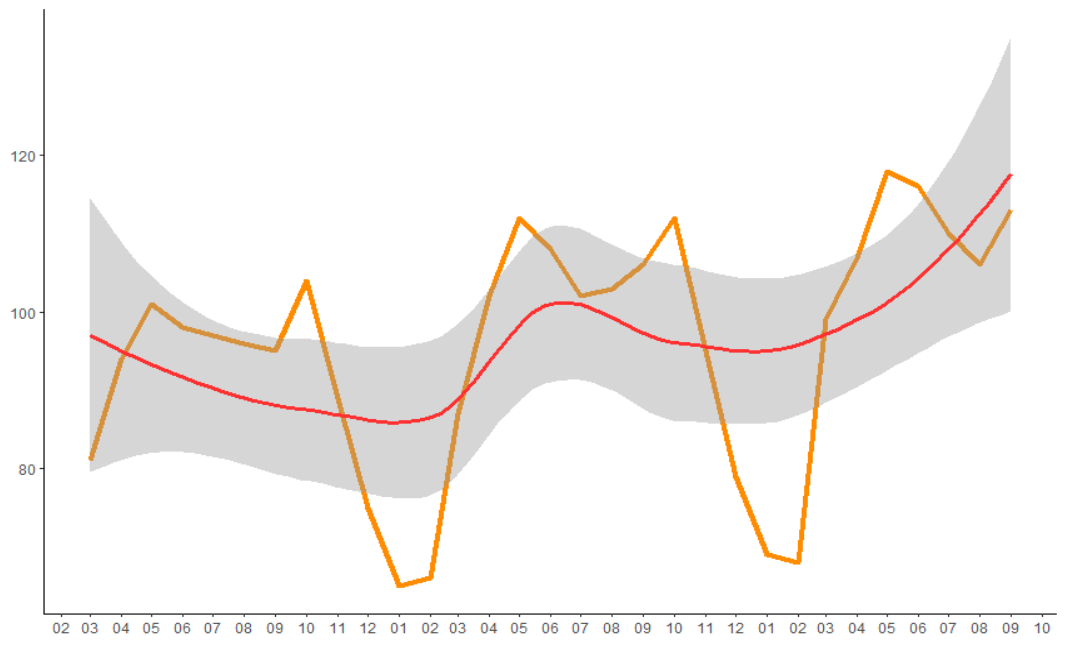


- 월별에 쓰인 카드 사용 내역(x는 월별, y는 백 억원 단위)(ggplot 함수 사용)

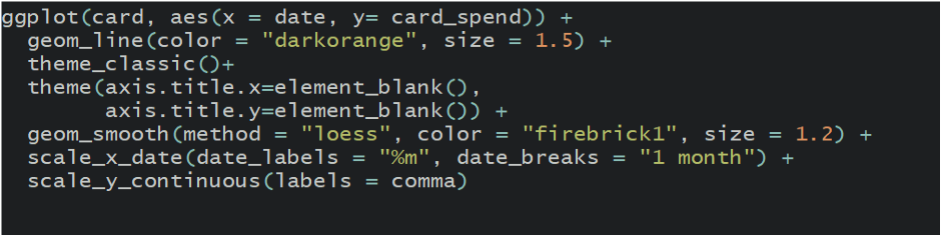
데이터 종류 : 월별 문화 공연 관람에 쓰인 카드 사용 내역

데이터 개수 : 31row(2016년 3월 부터 2018년 9월까지 월 단위 개수) \* 2column(날짜(월 단위), 카드 사용 금액(단위 백 억원))

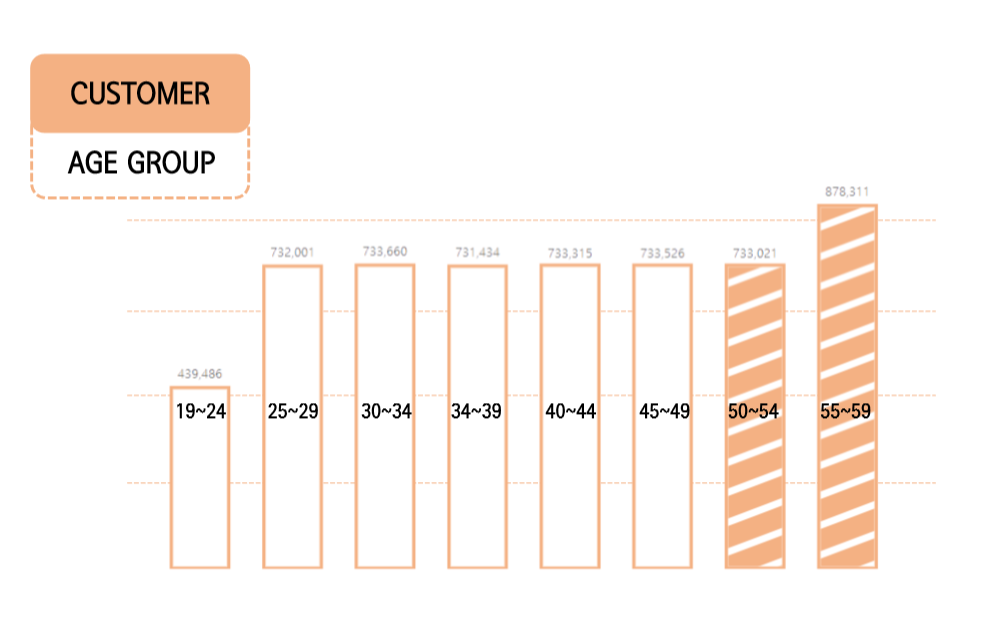
ggplot2패키지에 있는 geom\_line함수를 사용해서 문화 공연 관람에 쓰인 카드 사용 추세를 알 수 있다. x 값은 월별단위의 날짜이고, y 값은 문화 공연 관람에 쓰인 전체 카드 사용 금액이다. 단위는 백 억원이다. 추세선의 기울기가 증가하고 있는 걸로 봐서는 문화 공연 관람에 쓰인 금액이 증가하고 있다는 것을 알 수 있다.



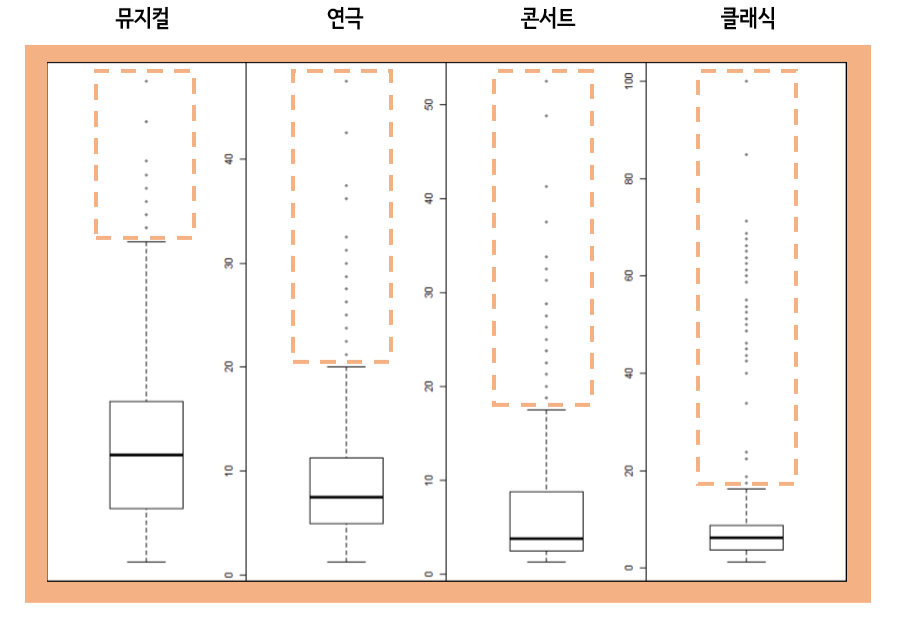
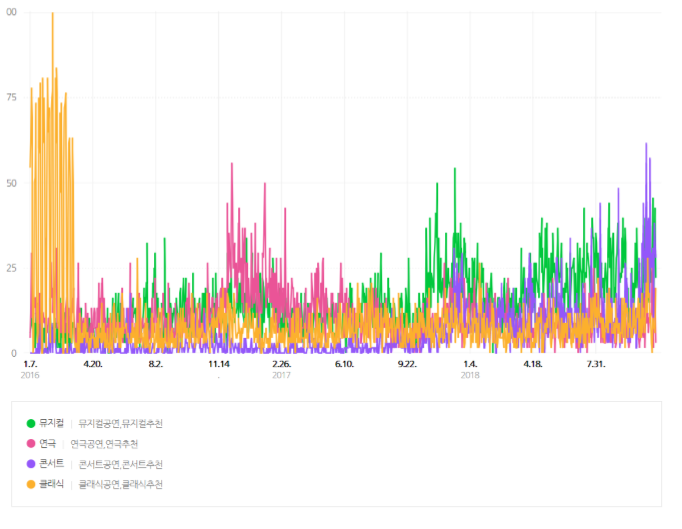
\* 코드



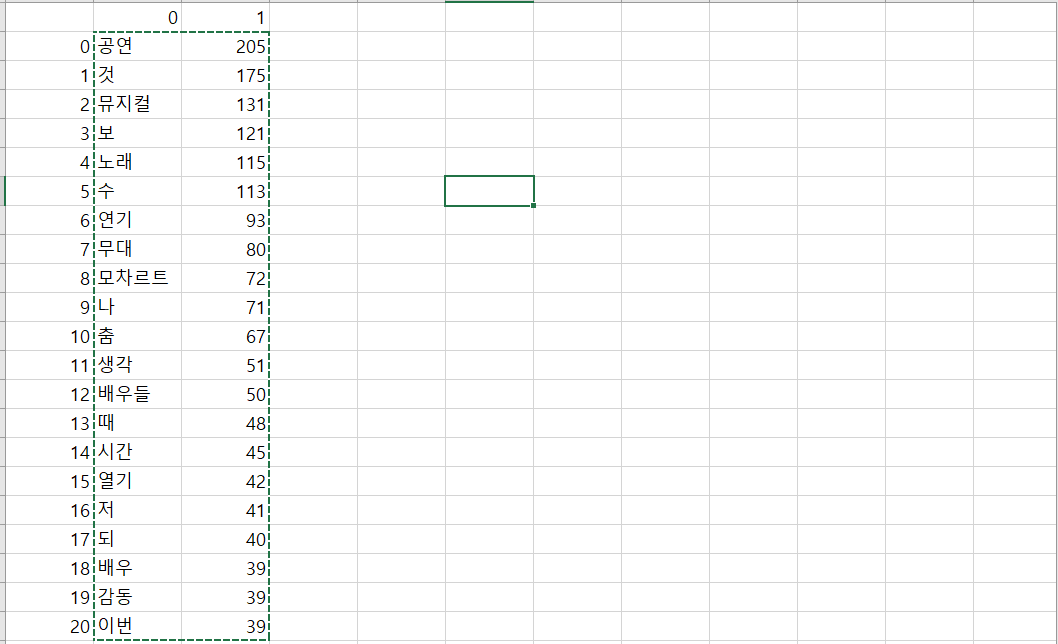
**(2) 관람객, 공연 주제 타겟팅 분석**



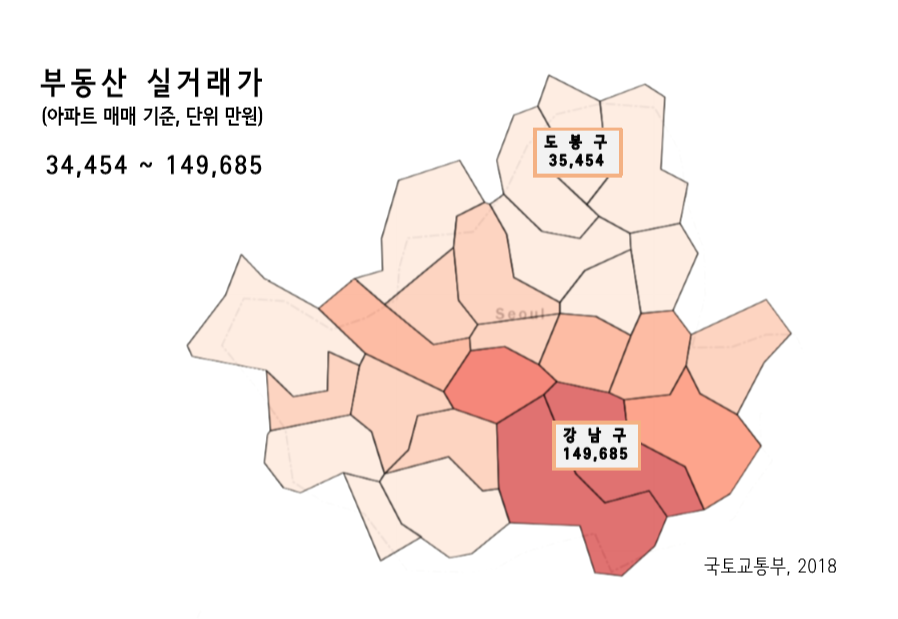
- 고객정보를 분석한 결과 의미 있는 결과값을 많이 얻을 수 없었다. 그래서 고객의 나이에 초점을 맞춰 가장 많은 비율을 가지고 있는 50대 회원을 대상으로 공연을 계획한다.



- 네이버 데이터 랩에서 가져온 결과에 따르면 50대의 뮤지컬, 연극, 콘서트, 클래식 검색 키워드의 선호는 클래식이 가장 높아 보이나, 분석결과 클래식은 공연 클래식 외에도 이더리움 클래식 등 다른 의미의 키워드가 많이 포함된 것으로 나타났고, 해당 이상치들을 처리한 결과 가장 검색수가 많았던 것은 뮤지컬로 판단이 되었다.



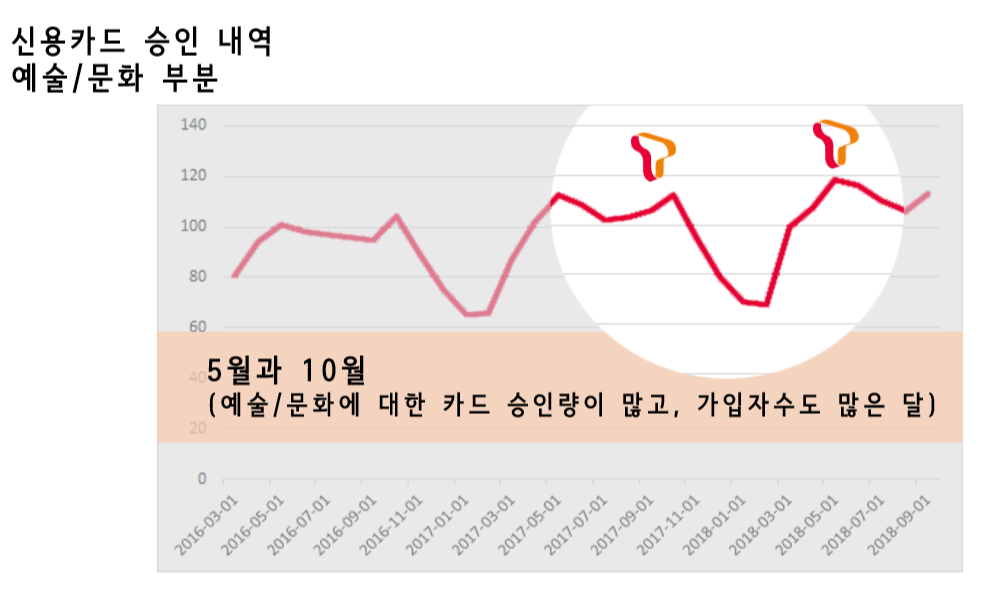
- 인터파크 리뷰를 처리한 결과 가장 많이 언급되는 의미 있는 키워드들은 모차르트, 감동, 토요일 밤, 임태경, 헤드윅, 토니, 박효신 등으로 나타났다. 해당 키워드들을 바탕으로 공연을 기획한다.



- 부동산 실거래가를 분석한 결과 강남구가 가장 높은 것으로 나타났고, 실제 거주자 수도 강남구 높은 것으로 나타났다. 이를 바탕으로 소득수준, 소비패턴을 예측하여 공연장의 위치는 강남구로 선정하였다.



- 티켓의 가격은 인터파크에서 크롤링한 데이터를 바탕으로 가장 많이 언급된 극장의 평균 티켓 가를 기준으로 산출하였다.



- 공연의 시기는 카드승인내역 분석의 결과 5월과 10월에 예술/문화에 대한 승인량이 높은 것으로 나타났다.

* **산출물(환경구성, 분석, 기타 코드)**
* Cloudera 환경 구성

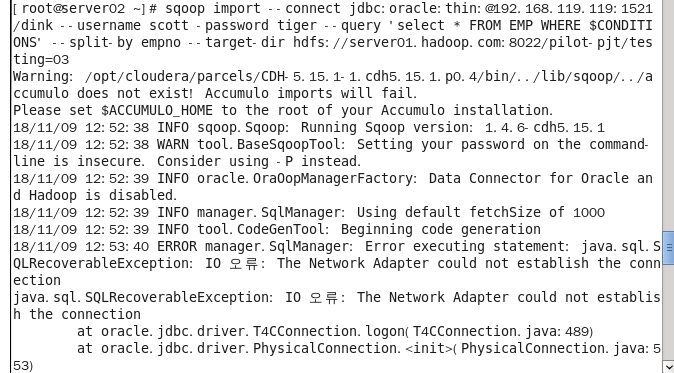


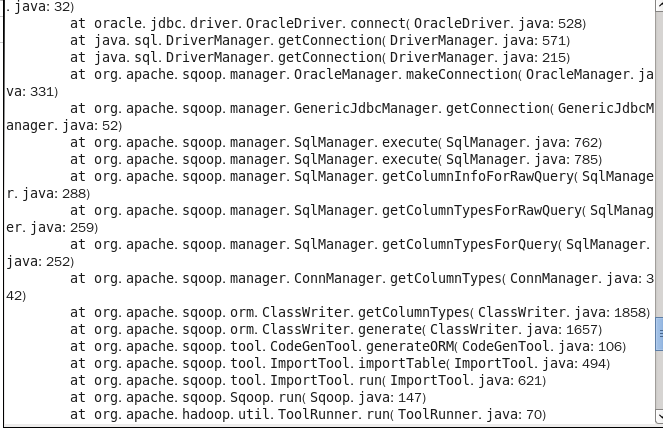
Server 01 : HDFS, YARN,

Server 02 : Hue, Sqoop 1 Client, YARN, Hive

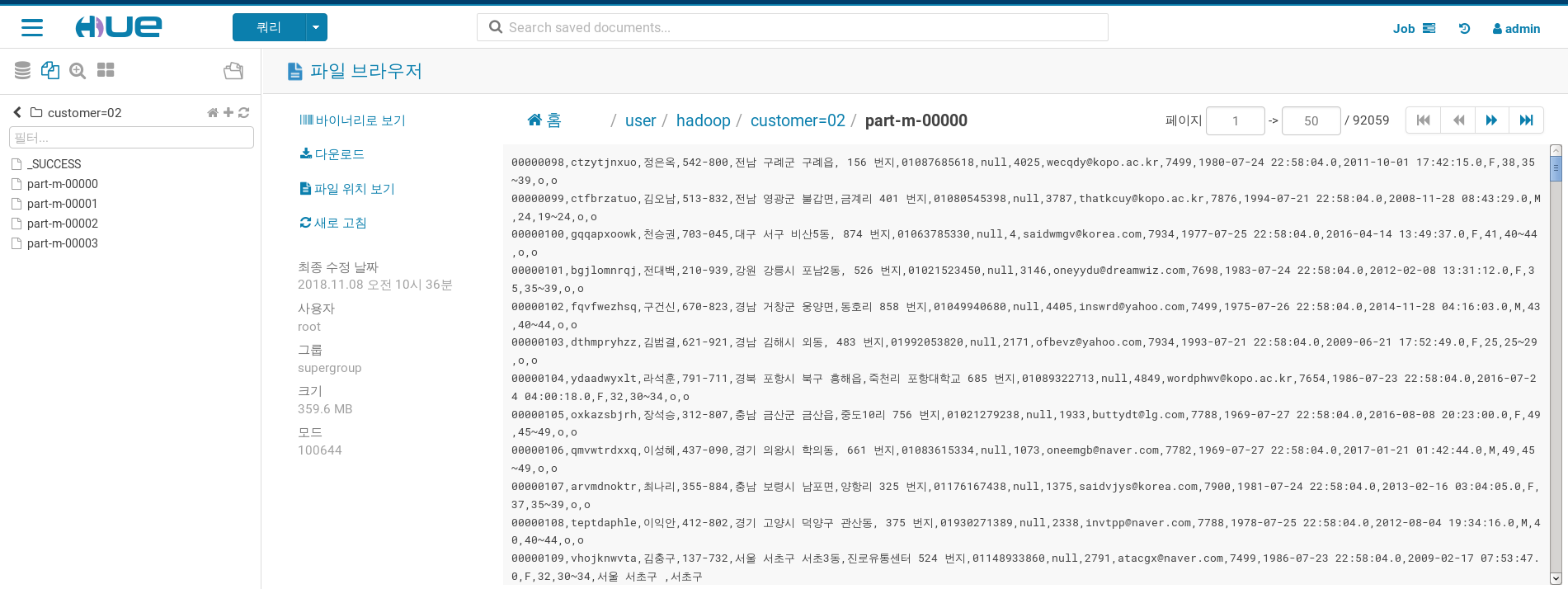
Server 03 : Cloudera

* Sqoop을 통해 Oracle DBMS 데이터 가져오기





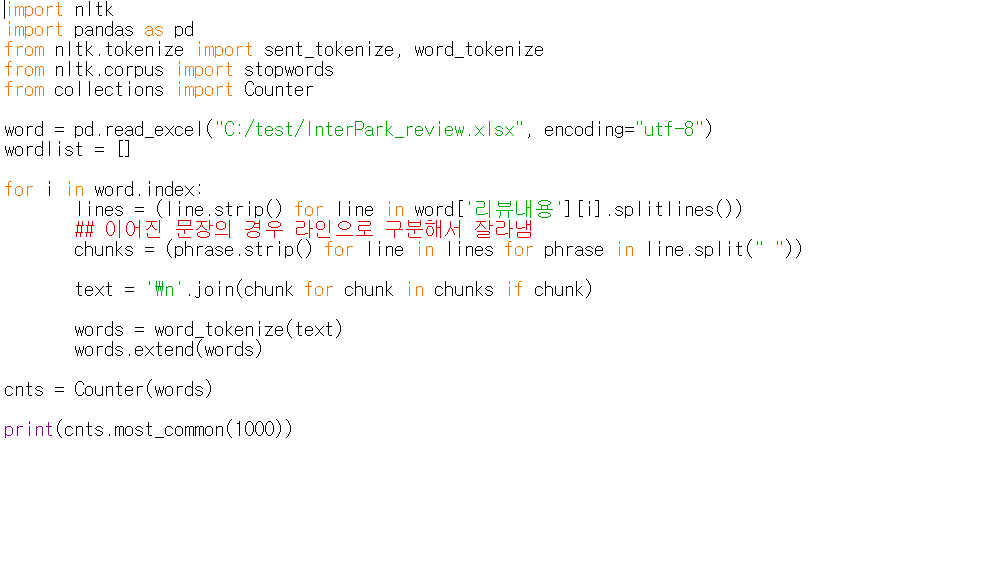
* Hue를 통해 CUSTOMER 데이터 관리 (Hue 위에서 Hive 사용 가능)



* 크롤링



* 자연어처리(인터파크 리뷰데이터 분석)



1. **결론**

고객 데이터 분석 결과 고객 데이터가 이미 Masking 되어 있는 데이터라 컬럼들의 의미 있는 관련성, 결과를 얻지 못하였다. 그래서 나이 정보에 집중해 분석할 결과 50대의 비중이 고객 데이터에서 가장 높다는 결과를 얻었다. 이를 바탕으로 네이버 데이터 랩에서 50대가 가장 많이 검색하는 공연의 종류는 뮤지컬이라는 결과를 얻었고, 공연의 카테고리를 뮤지컬로 잡게 되었다. 인터파크 리뷰 중 가장 공감도가 높았던 데이터를 크롤링해서 분석한 결과 특정 배우의 이름이 많이 거론되는 것을 알 수 있었고, 이를 바탕으로 공연을 기획한다면 좋은 반응을 얻을 수 있을 것으로 기대된다. 예산의 경우 리뷰데이터에서 수집한 공연장정보를 이용해 많이 거론된 특정 공연장들의 티켓 값의 평균값을 티켓 값으로 잡고 계획한다.

1. **제한사항**

- CUSTOMER table이 Masking된 데이터라 분석결과로 정확한 값을 얻을 수 없었음

- 국내에서 공연되고 있거나 공연될 뮤지컬들의 모든 데이터를 얻는데 한계가 있음

- 경제력 / 문화생활 선호 데이터의 수집이 한계가 있음

- HDFS 시스템을 사용하기에 하드웨어의 사양이 너무 낮음

- 문화/취향에 대한 분석 방법론 고찰 필요

- HDFS 시스템을 구성하던 중 각 가상서버끼리의 연결이 불안정하여 Oracle DB에서 데이터를 가져와 입력하는 중에 커넥션 에러 발생 > 노트북의 가상서버끼리의 연결 문제로 보임

- RStudio에서 Oracle DBMS로의 커넥션 문제 : DBMS의 리스너 문제

- RStudio에서 DBMS 의 데이터들을 수정하면서 DBMS가 있는 가상 서버의 디스크용량이 다 참 > 부팅이 안되는 문제

- 크롤링을 할 때 특정 ip를 차단하거나 보안으로 인해 html 문서를 받아 올 수 없었던 문제, 웹사이트의 태그 구조상 크롤링이 안되는 경우도 있었음

1. **보완사항**

* 정확한 Customer Data로 분석
* 뮤지컬에 대한 정성적/정량적 Data 확보
* 빅데이터 처리가 가능한 고사양 하드웨어 확보