**빅데이터 보고서**

작성자 : 박성준

**빅데이터 주요 업무 세부사항**

1. **배달데이터 시각화 및 인사이트 도출**

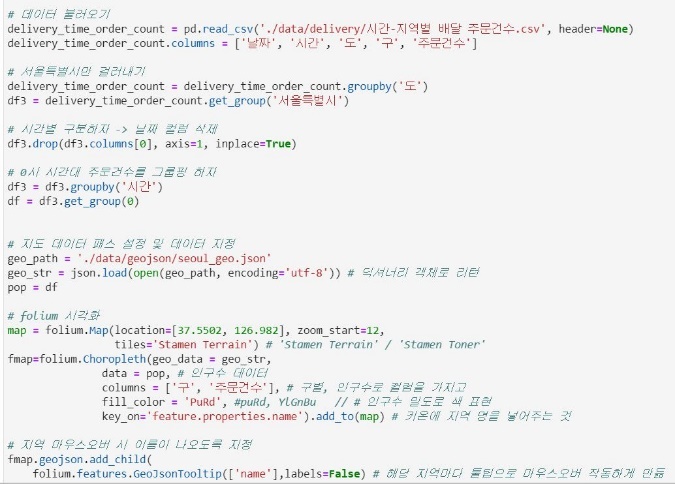
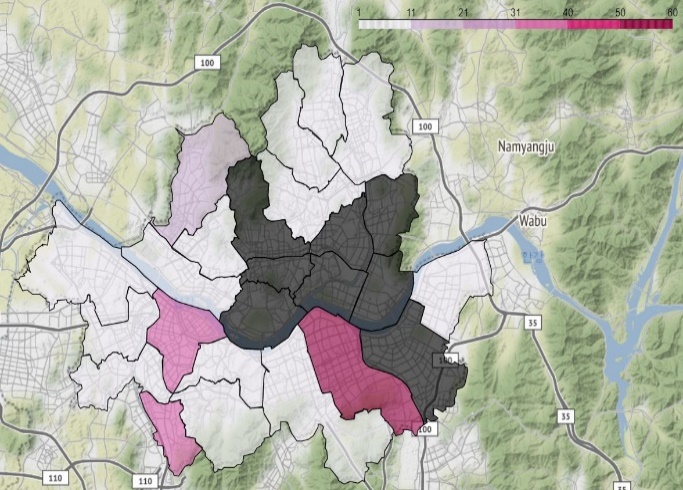
* KT 빅데이터 플랫폼과 한국데이터거래소 등의 배달 관련 데이터를 수집하고 이를 확인하고자 하였습니다. 배달데이터는 전국 단위로 있었기에 서울시 단위로 그룹핑하였습니다. 이러한 데이터를 시각화한 결과 특정 구에서는 데이터가 존재하지 않으며, 비교적 강남구에 배달데이터가 많다는 것을 알 수 있었습니다. 때문에 예상주문건수, 평균배달시간, 평균주문금액 등을 **강남구**에 초점을 맞추어 찾아보는 작업을 수행하였습니다.



<배달데이터 확인>



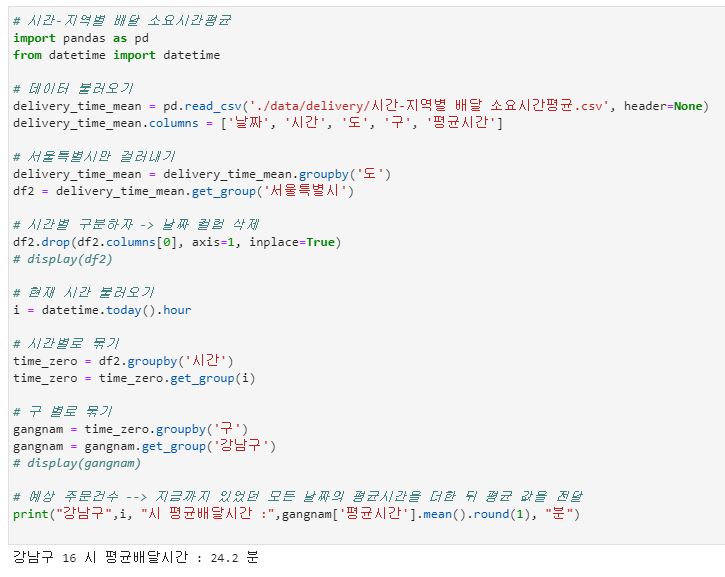
<배달데이터 확인>

<시간-지역별 배달주문건수 지도시각화>



<강남구 실시간 예상주문건수 파악>



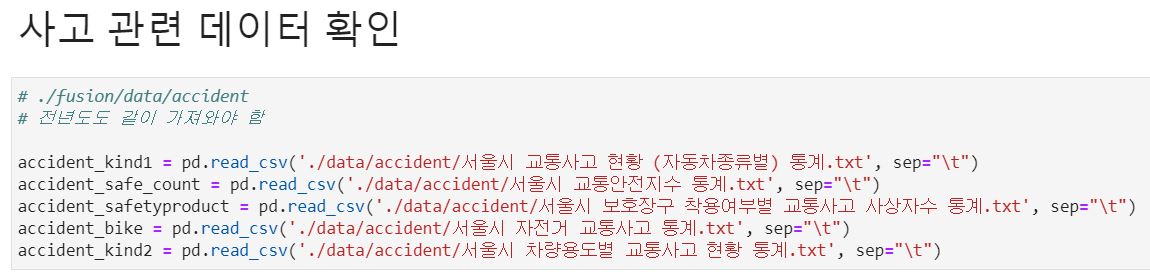
<강남구 실시간 평균배달시간 파악>



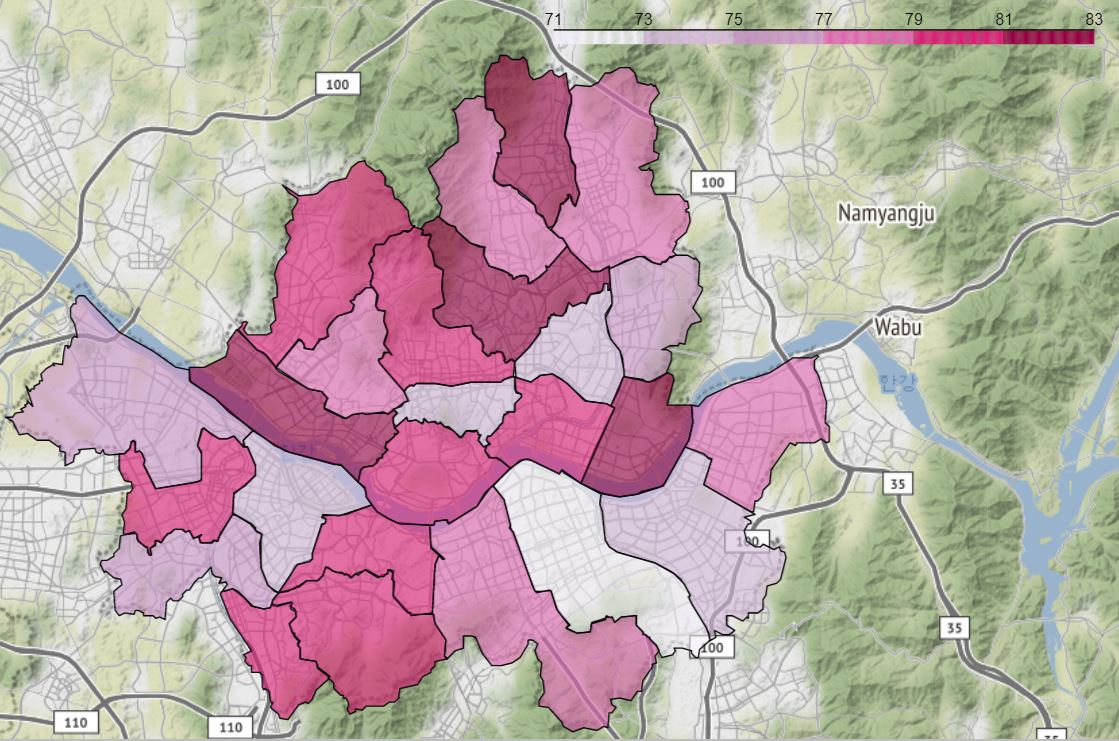
<강남구 실시간 평균주문금액 파악>

1. **안전사고 데이터 전처리 및 시각화를 통한 인사이트 도출**

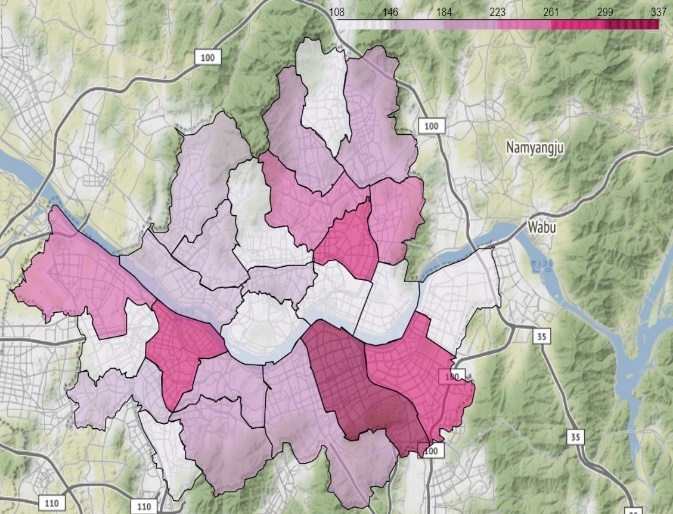
* 공공데이터포털, 그리고 서울 열린데이터광장 등에서 배달 및 오토바이 사고에 대한 분석과 시각화를 진행해보았습니다. 이를 통해 서울 내에서 강남구의 교통안전지수가 가장 낮고, 오토바이 사고 건수가 가장 많다는 것을 알 수 있었습니다. 때문에 앞으로 서비스 구현의 예상지역을 **강남구**로 설정하였습니다.



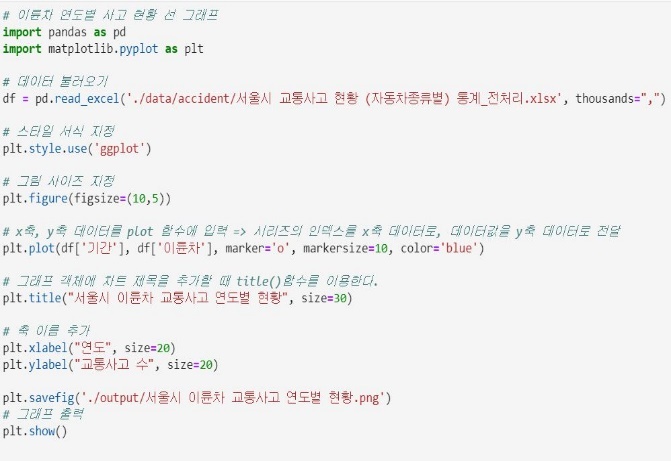
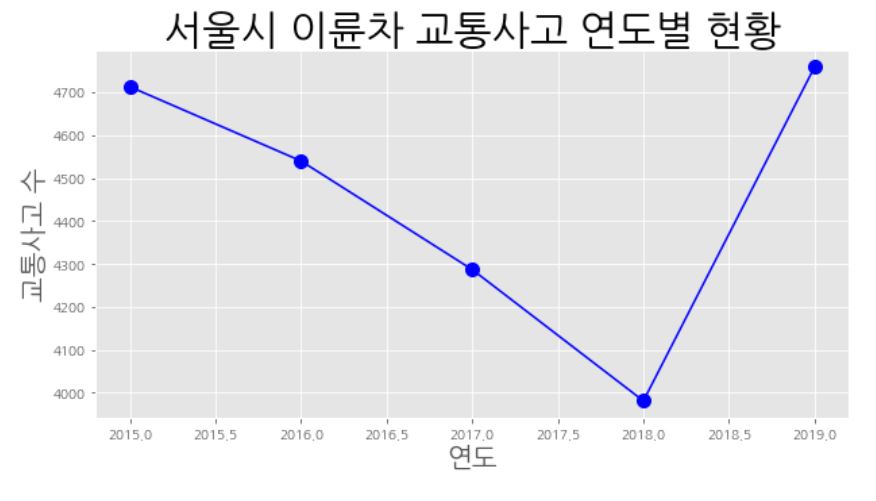
<사고 관련 데이터 확인>

<서울시 교통안전지수 통계 시각화>

<서울시 차량용도별(이륜차) 교통사고 시각화>

<서울시 이륜차 교통사고 연도별 현황>

1. **회귀분석을 통한 상관관계 파악 -> 없음**

* 앞서 확인했던 배달데이터 중 서울을 중심으로 배달데이터를 회귀분석하여 주문건수에 영향을 주는 요인을 알아보고자 회귀분석을 시도하였습니다. 회귀분석을 위해 다음과 같은 순서를 거쳤습니다.

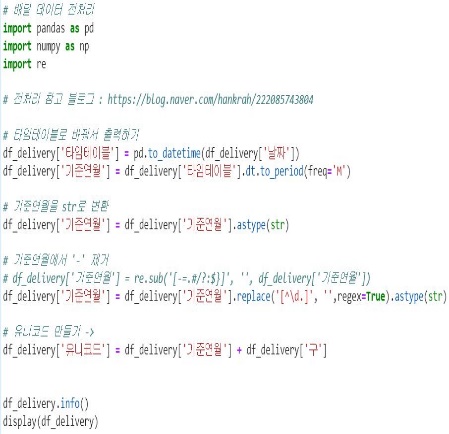
1. 데이터 확인

<데이터 확인>

1. 데이터 병합을 위해 키 데이터를 생성한 후, on 옵션을 이용해 데이터 병합

* 키 데이터를 생성할 때, ‘연월’ + ‘구’ 를 합쳐 유니크한 데이터컬럼 생성



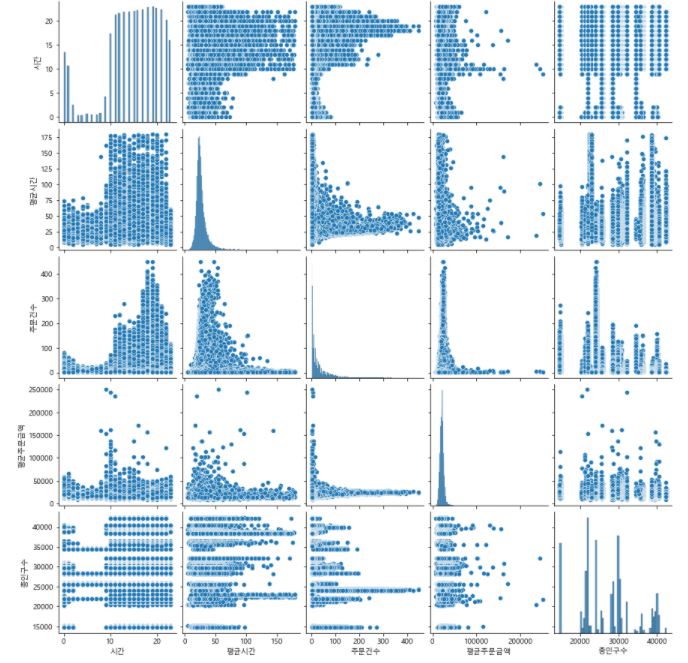
<데이터 유니크코드 생성 및 통합>

1. 회귀분석을 위한 데이터 선별



<필요없는 컬럼 제거>

1. 시각화

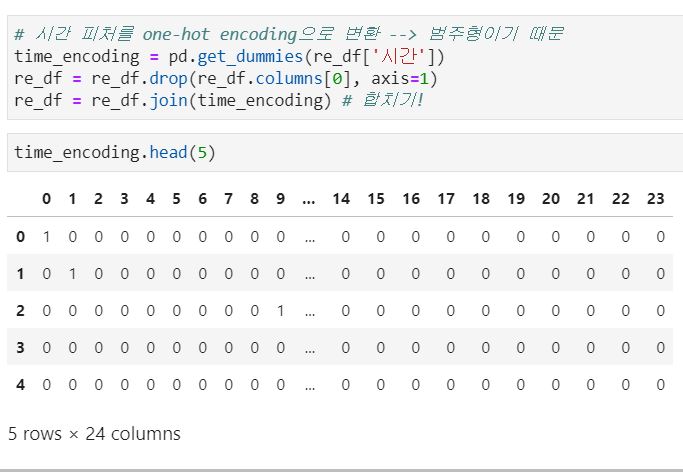
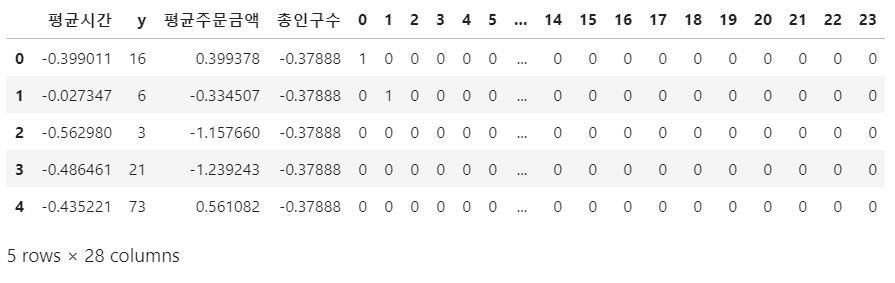


<모든 변수 간의 시각화>

1. 스케일링 및 구간화

<표준화>

<시간 구간화>

1. 결과 확인



<회귀분석 결과 확인>

**R스퀘어값을 확인**한 결과, **0.25**로 종속변수(예상주문건수)와 독립변수(다른 변수) 사이에는 **큰 상관관계가 없다**는 것을 알 수 있었습니다. 때문에 전진선택법이나 후진제거법, 단계적 선택법을 이용하더라도 크게 의미있는 상관관계를 얻을 수 없을 것이라 판단하여, 이후 과정은 진행하지 않았습니다.

1. 연관분석을 통한 배달 및 안전의 연관성 분석

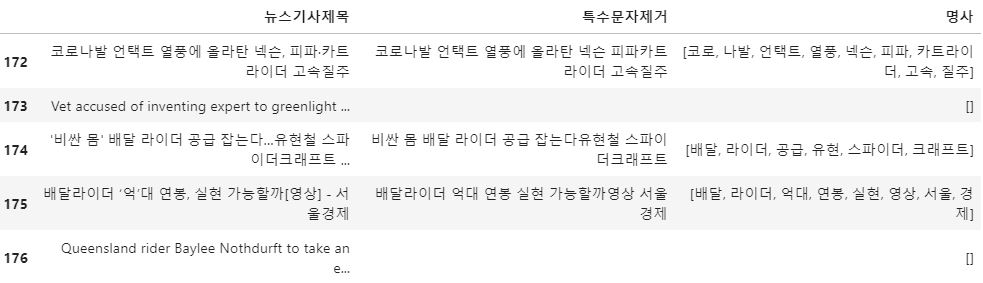
* 배달라이더와 안전이 상관이 있는지에 대해 알아보고자 구글에서 키워드 ‘라이더’를 입력하여 뉴스 기사를 크롤링하였습니다. 이를 기반으로 명사를 추출하였고, 연관분석을 통해 ‘라이더’와 어떠한 단어가 연관이 있는지 알아보고자 하였습니다. 실제 연관분석을 수행한 결과 **‘라이더’, ‘배달’, ‘안전’ 등이 많은 연관성**을 가지는 것이 확인되었습니다.

1. 뉴스기사 크롤링

<뉴스기사 크롤링>

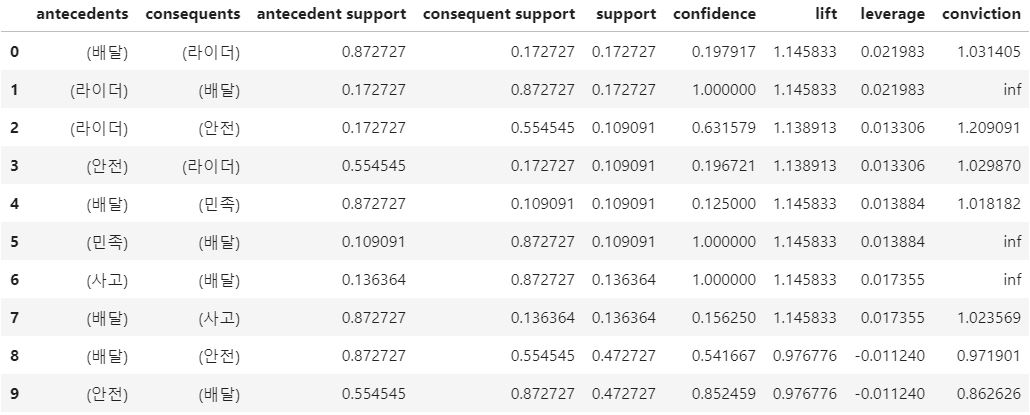
 



<특수문자 전처리 및 명사 추출>

1. 라이더 키워드 연관분석



<라이더 키워드 연관성 분석>

해당 표에서 antecedents에 해당하는 키워드가 포함된 글에서 cosequents에 해당하는 키워드를 포함하는 글의 확률이 confidence입니다. confidence만을 보고 판단하였을 때, 라이더라는 키워드를 검색하였을 때 안전과 관련된 단어가 많으며 높은 상관관계를 가진다는 것을 파악할 수 있었습니다.

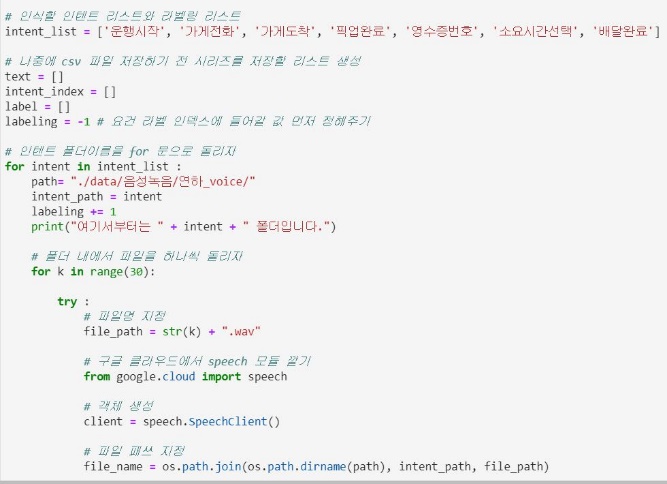
1. STT 인텐트 분석 기본 구조 설계

* STT나 TTS를 이루는 기술 자체를 개발하는 것은 힘들었기에, 중간과정에서 처리되는 데이터를 어떻게 분석할 지에 대해 고민해보았습니다. 기본적으로 배달운행에 사용되는 발화내용은 한정적이라 예상했습니다. 때문에 특정 목적을 갖는 발화내용을 예상해 녹음하여 이를 STT로 변환한 후, 이를 형태소 분석하여 해당 intent을 분석하고자 하였습니다. 이러한 기본 마인드에서 더 나아가 AI가 각 발화내용별 고도화를 진행하였습니다.

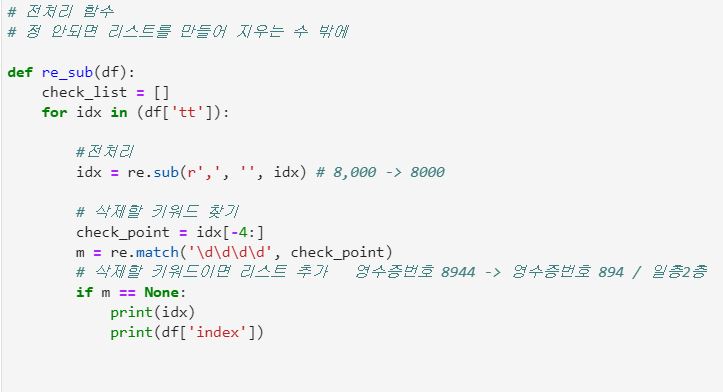


1. AI 학습을 위한 STT API활용 학습데이터 구축

* AI에서 학습할 수 있는 데이터를 만들기 위해 어플리케이션 사용 시 등장할 발화내용을 녹음하였습니다. 이를 이용하여 STT처리를 통해 텍스트로 변환하고, 변환된 텍스트를 학습에 맞춰 전처리하였습니다. 그 과정 중 intent에 맞춰 라벨링을 주었습니다.

<STT변환 및 데이터 저장>

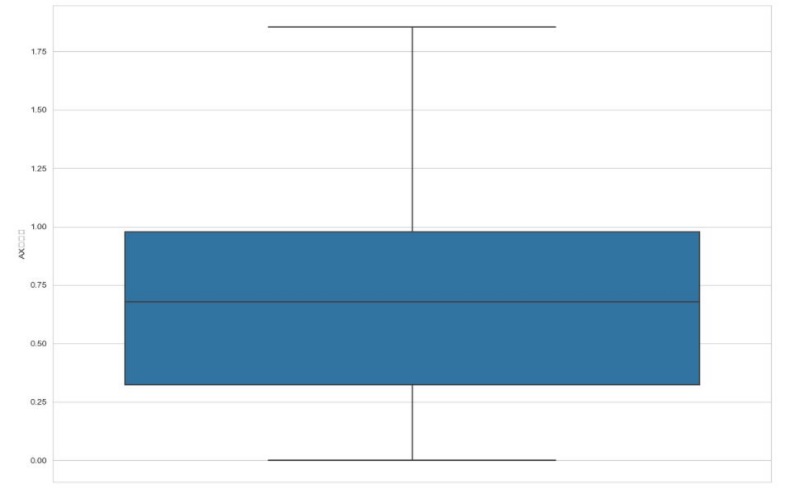
<AI학습을 위한 데이터 전처리>

1. 자이로센서 예상데이터 실시간 이상치 설계 및 군집분석 시도 / 실제데이터 분석

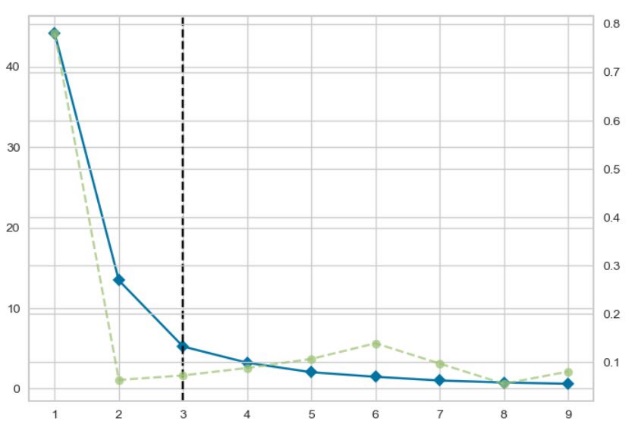
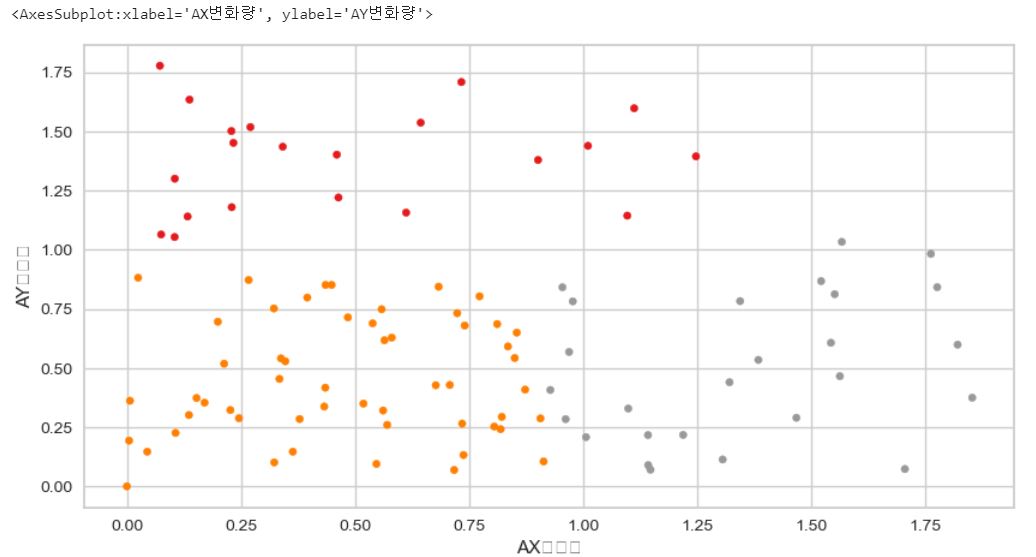
* 프로젝트 중 자이로센서는 프로젝트 마감 2주전에 도착하였습니다. 그렇기에 보험데이터 설계를 위한 자이로센서 데이터가 없었습니다. 때문에 사전에 예상데이터를 난수로 만들어 실시간 이상치를 어떻게 감지할 지, 그리고 이러한 데이터를 어떻게 분석할 지에 대한 설계를 진행하였습니다. 이때 난수는 정규화한 상태를 가정하여 -1에서 1사이의 값을 준 뒤, 직전값과 현재값의 차를 변화량으로 설정하였습니다. 이를 바탕으로 군집분석을 한 결과, 깔끔하게 군집이 되는 것을 알 수 있습니다.

<예상데이터 난수 컬럼 생성>

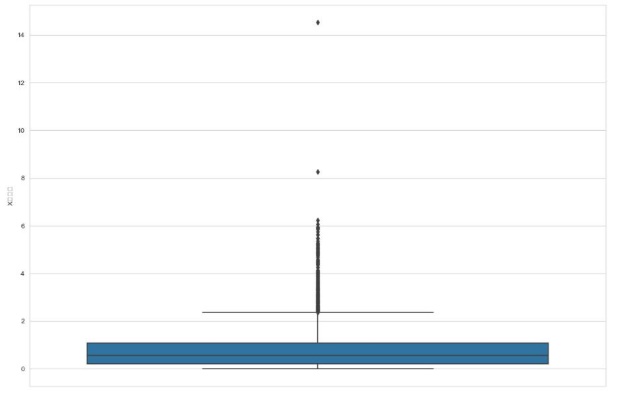
 

<박스 플롯을 통한 이상치 확인>

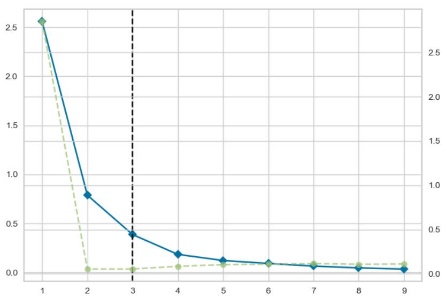
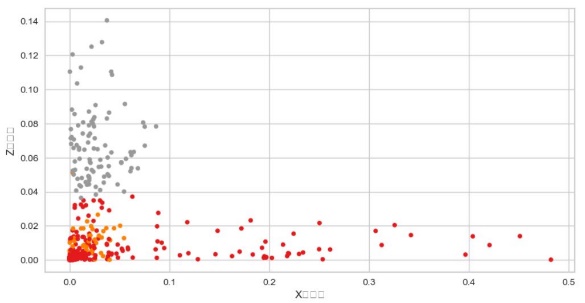
 

<군집 내 변동성 파악 및 군집분석>

* 하지만 실제데이터를 기반으로 이상치를 측정하고 군집분석을 한 결과, 깔끔하게 구별하는 것은 어려웠습니다. 때문에 비지도분석을 통한 분류보다 지도분석을 통한 분류가 필요할 것으로 예상됩니다.

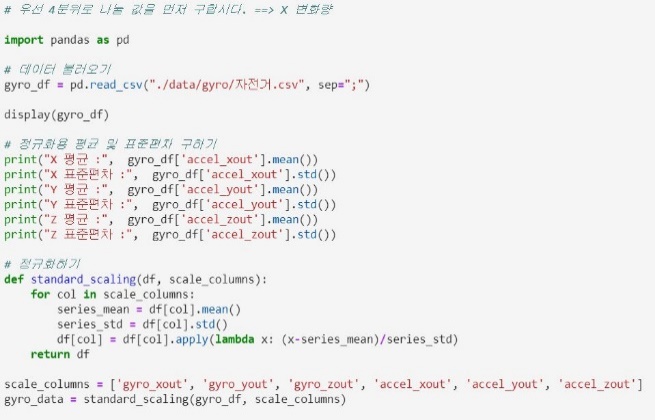
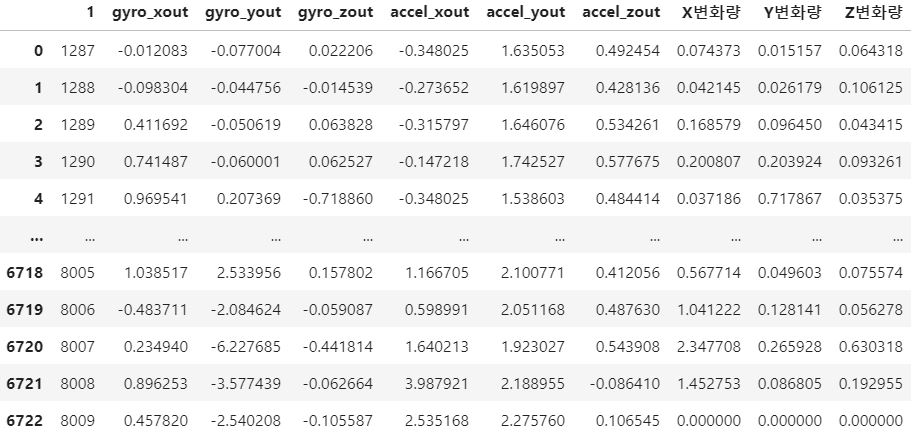
<박스플롯을 통한 이상치 확인>

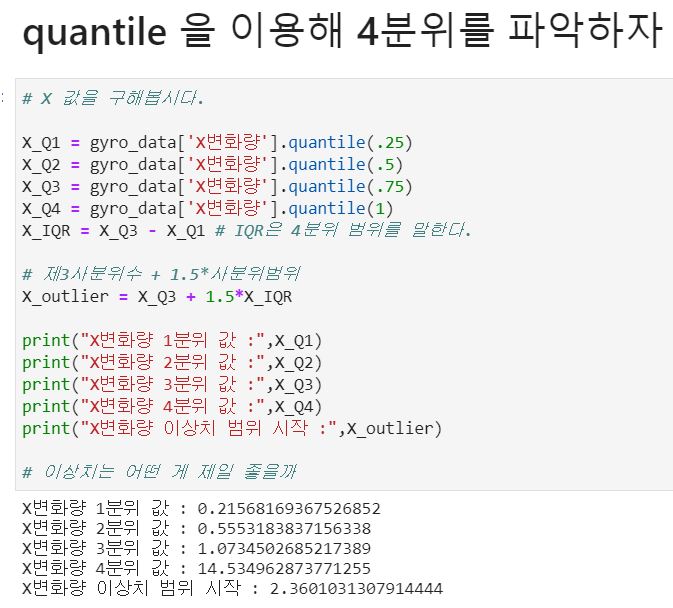
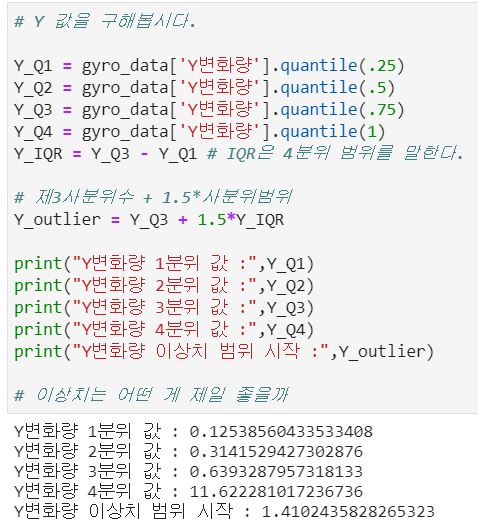
<군집 내 변동성 파악 및 군집분석>

1. 보험(멘토링) / 이상치 감지 구조 설계 및 라즈베리 파이 내 구축

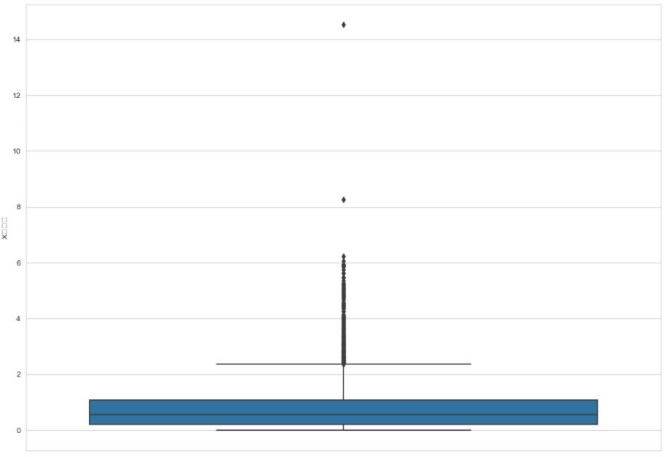
* 멘토님께 보험에 대한 조언을 듣고, 어떻게 보험이라는 영역을 배달 라이더 서비스에 구현할 수 있을지 고민해보았습니다. 고민한 결과, 자이로센서의 X축과 Y축이 주행에 많은 영향을 끼친다는 사실을 파악하여 해당 값을 주요변수로 설정하여 이상치를 찾아내도록 구상하였습니다. 예컨대 실시간으로 데이터가 쌓이는 **자이로센서 데이터의 직전값과 현재값의 차이가 이상치 범위에 해당할 경우 alert 처리**할 수 있도록 만들었습니다.

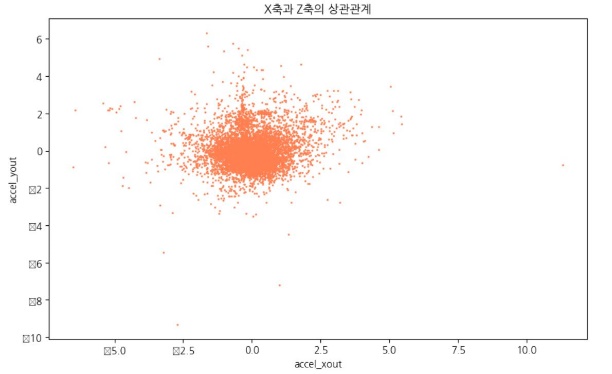
<자이로센서 데이터 평균 및 표준편차 추출 / 정규화 / 변화량 컬럼 생성>

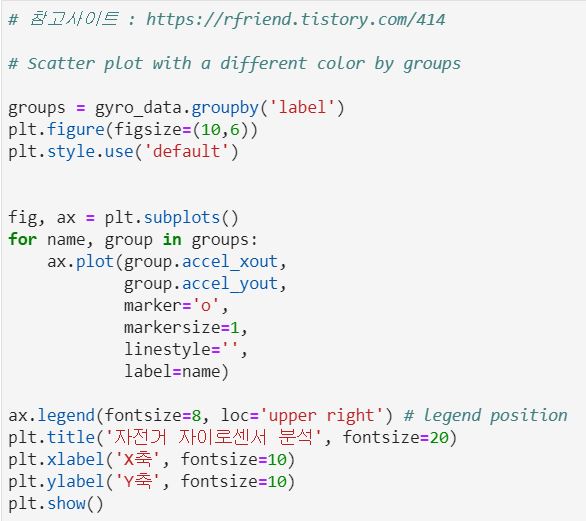
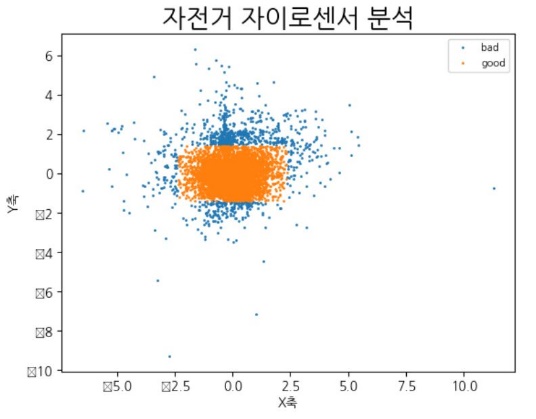
<자이로센서 X,Y 축 변화량의 4분위 값>

<박스플롯을 통한 X변화량 이상치 파악>

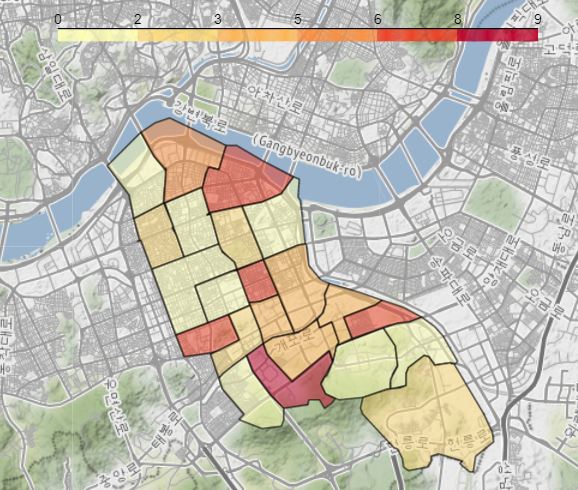
<산점도를 통한 데이터 분포 파악>

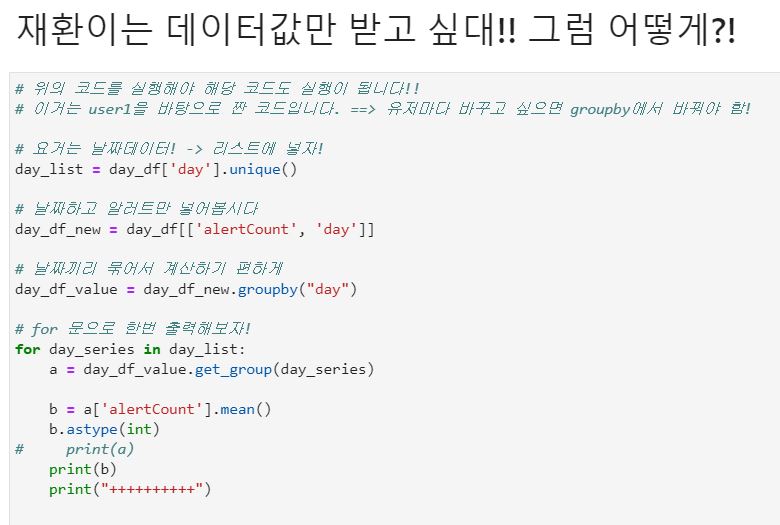
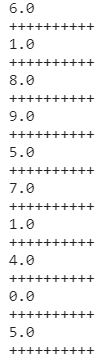
<이상치를 기준으로 데이터 분류분석 시각화>

1. 보험데이터를 통한 모니터링 페이지 시각화

* 누적된 보험데이터를 실시간으로 파악하고 관리하기 위해 가상 데이터를 만들어 시각화를 진행하였습니다. 예컨대 유저의 일일 위험감지 횟수, 특정지역 위험주행횟수 등을 시각화하였습니다. 이러한 **시각화는 클라우드가 진행한 모니터링 페이지에서 확인할 수 있도록 제공**하였습니다.

<강남구 동별 위험주행횟수 시각화>

<특정 유저의 날짜별 위험주행횟수>

**느낀 점 및 보완할 점**

**느낀 점**

* 처음 융복합 프로젝트를 진행할 때는 많은 걱정이 있었습니다. AI는 빅데이터와 관련이 있는 부분이 있어 어느 정도 파악할 수 있었지만, IoT나 Cloud와 같은 경우에는 어떠한 사전지식도 없었기 때문입니다. 하지만 프로젝트를 진행하면서 각 분야에 대한 설명을 팀원에게 들을 수 있었고, 해당 분야에 대한 인지범위를 넓힐 수 있었습니다. 이번 프로젝트를 진행하며 단순히 IT역량 뿐만 아니라 커뮤니케이션 역량 등을 제고하며 많은 것을 배웠습니다.

**보완할 점**

1. **회귀분석을 위한 데이터수집 프로세스를 구축해야 한다.**

* 회귀분석을 진행하며 부족한 예상주문건수를 예측하기 위해서는 더 많은 변수가 필요하다고 생각하였습니다. 때문에 보험데이터를 저장하는 과정에서 예측에 필요한 데이터를 추가적으로 지정한다면, 이후 더욱 효과적인 데이터 분석을 이룰 수 있으리라 판단됩니다.

1. **이상치 분석을 세분화하여 위험등급 정도를 나누어보아야 한다.**

* 이상치 분석은 단순히 위험감지만을 잡아낼 수 있도록 설계하였습니다. 때문에 이를 U턴위험감지, 급회전 위험감지, 급제동 위험감지 등과 같이 특정 위험감지 분류를 이룰 수 있다면, 보험설계 측면에서 더욱 효과적일 것이라 판단됩니다. 또한 특정 주기를 반복하거나 지나치게 안정적인 치트데이터를 찾아내는 과정을 수행한다면 보험료 책정 부분에 도움이 될 것이라 생각합니다.

1. **비지도학습(군집분석)으로 분류가 되지 않는다면, 각 유형별(유턴, 회전 등) 규칙을 찾아내고 AI가 학습할 수 있도록 라벨링을 해야한다.**

* 군집분석으로 데이터를 분류하려 시도하였지만, 적절한 분류를 이루지는 못했습니다. 때문에 각 유형별 라벨링 등을 통해 지도학습으로 U턴 혹은 급제동 등을 찾아낼 수 있다면, 자이로센서 분석을 좀 더 고도화시킬 수 있으리라 생각합니다.

**데이터 목록**

* **배달데이터(KT 빅데이터 플랫폼) :**

1. 배달상점 데이터 (42769 rows × 11 columns)
2. 시간지역별 배달 소요시간 평균 (375220 rows × 5 columns)
3. 시간 지역별 배달 주문건수 (381068 rows × 5 columns)
4. 시간 지역별 배달 평균주문금액 (374561 rows × 5 columns)
5. 업종 지역별 배달 주문건수 (197192 rows × 5 columns)
6. 업종 지역별 평균배달소요시간 (195986 rows × 5 columns)
7. 업종 지역별 평균주문금액(191306 rows × 5 columns)
8. 주문지역 인구 특성 (157374 rows × 8 columns)

* **연관분석 :** 구글 기사 크롤링 (185개)
* **AI학습데이터 :** 녹음파일 (660개)
* **보험데이터 :** 자이로센서 (자전거 – 6723 rows × 7 columns)

**참고사이트**

1. **데이터 전처리**

* <https://blog.naver.com/hankrah/222085743804>

1. **텍스트 연관분석**

* <https://needjarvis.tistory.com/59>
* <https://lemontia.tistory.com/903>

1. **군집분석**

* <https://tariat.tistory.com/819>
* <https://m.blog.naver.com/samsjang/221017639342>

1. **이상치 감지**

* <http://www.databaser.net/moniwiki/wiki.php/%EC%9D%B4%EC%83%81%EC%B9%98%EC%A0%9C%EA%B1%B0%EB%B0%A9%EB%B2%95>
* <https://sosomemo.tistory.com/34>

1. **지도시각화**

* <https://blog.naver.com/kcchang61/221350672356>