자료분석 보고서

이름: 유성민

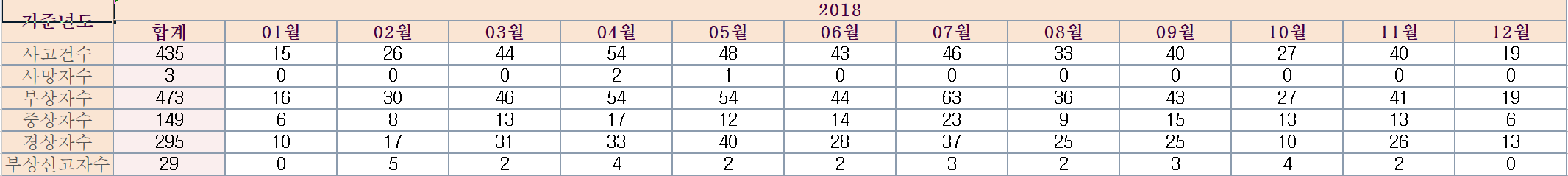
학번: 201520908

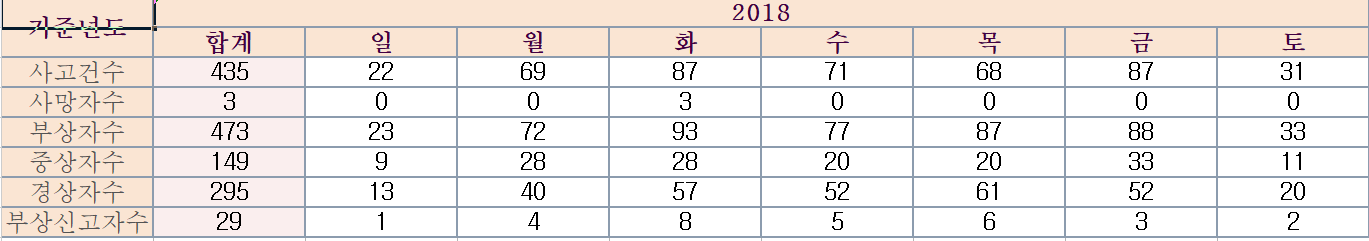
학과: 소프트웨어

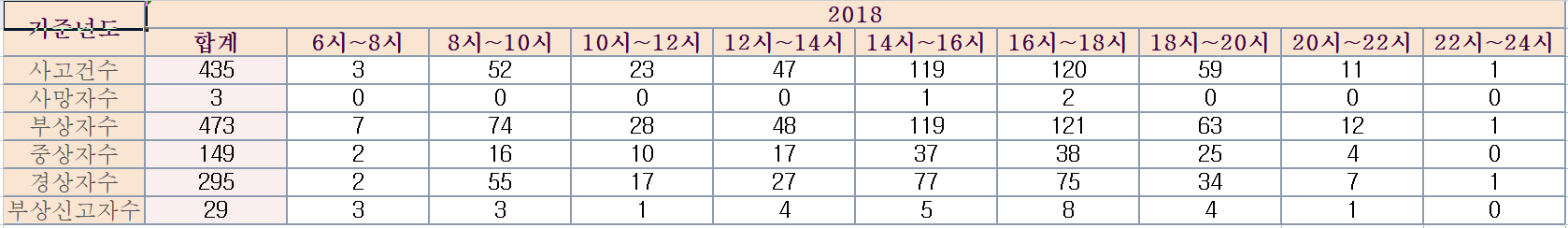
학년: 3

1번째 데이터: 도로교통공단의 교통사고분석시스템 어린이 교통사고 통계

<http://taas.koroad.or.kr/sta/acs/exs/typical.do?menuId=WEB_KMP_OVT_UAS_UDS>

전자 과실로 스쿨 존 내부의 어린이가 사망하였을 경우 3년 이상의 징역을 부과하는 내용이다. 하지만 민식이법은 제정 과정 중 드러난 피해자 유가족들의 거짓 증언과 가해자를 향 여론몰이로 인해 책임회피, 문재인 정부가 충분한 공론화와 검증 과정을 거치지 않고 2020년 3월을 목표로 섣불리 입법을 강행한 것에 따른 졸속 처리 논란, 행위에 대한 과잉처벌과 혹형 논란 등으로 국민의 실질적인 삶과 동떨어진 비현실적인 악법으로 비판을 받고 있다. 또한 원인 제공자 및 고의성 여부를 전혀 고려하지 않고 오로지 결과만으로 판단하는 것은 현대적 법리에 어긋나기 때문에 위헌의 소지도 있어 범국민적으로 찬반여론이 무척 거세게 일고 있다….(구글 검색 중 일부 발췌) 이 내용은 요즘 핫 한 이슈가 되고있는 민식이법입니다. 요즘 커뮤니티 사이트를 돌아다니다 보면, 스쿨 존을 피해서 돌아서 가야한다, 아예 운전대를 잡지 말아야 된다. 등등 운전자들의 걱정이 이만저만이 아니고 저도 운전자로써 굉장히 걱정되는 부분입니다. 따라서 저는 스쿨 존에서 발생한 월 별, 요일 별, 시간 별, 사고유형에 대해서 조사를 해보았습니다. 먼저 월 별입니다.

4월과 5월에 가장 많은 교통사고가 발생하였고, 당연한 것이겠지만, 학기가 방학보다 교통사고율이 높았습니다. 사망자 수는 부상자에 비해 많지는 않았습니다. 다음으로는 요일 별입니다.

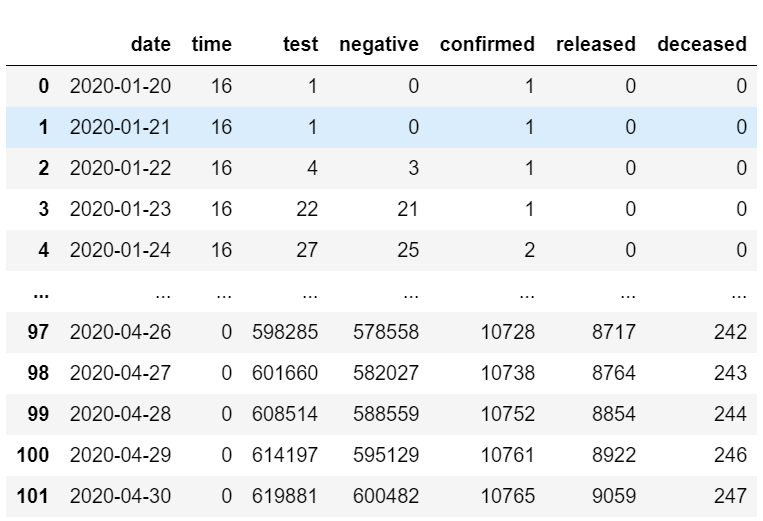
대다수의 사고가 주중에 발생하였고, 주말에는 감소하는 추세였습니다. 아무래도 애들이 학교를 안 가고, 놀러가거나 집에 있기에 그런 것으로 추정됩니다. 다음으로는 시간별 데이터입니다.

등교시간과 하교시간인 8시 ~ 10시와 하교시간 14~18시에 거의 대부분의 사고가 발생했습니다. 주로 하교시간에 많은 사고가 발생하는 것으로 미뤄볼 때, 등교는 아무래도 부모님과 함께 가는 경우가 많고, 하교는 대부분 아이 스스로 하는 경우가 대다수 이기 때문인 것으로 생각됩니다. 이제는 법규위반을 살펴보아 어떤 이유로 사고가 나는 건지에 대해 알아봤습니다.

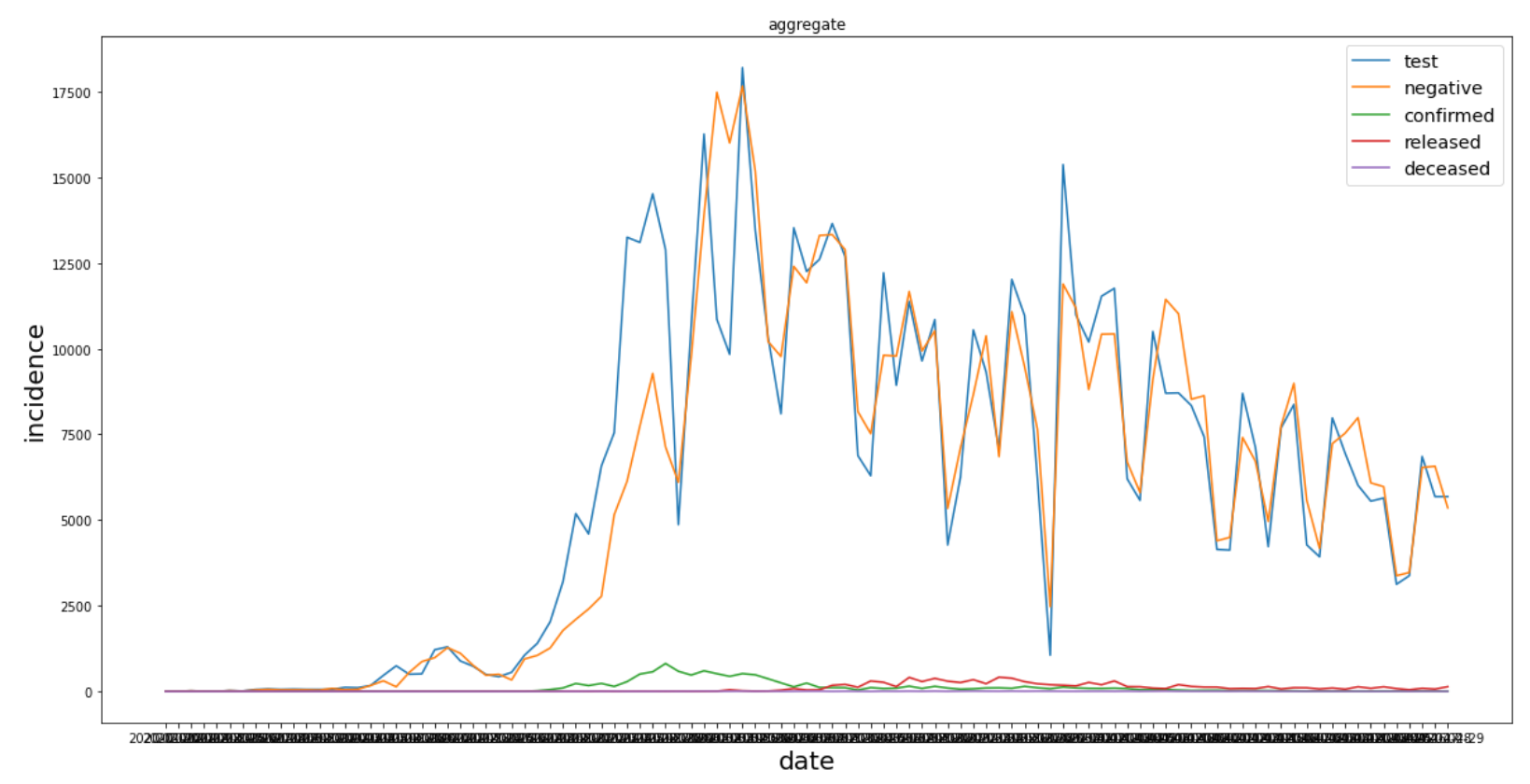
현 데이터에서 아쉬운 점은 교통공단에서 운전자의 과실만 취급한다는 것입니다. 유튜브나 기타 커뮤니티 사이트의 대부분의 영상들은 어린이들이 갑자기 튀어나와서 사고가 나는 영상들이 많고, 이것 때문에 민식이 법이 이슈가 되는 것인데, 그런 것에 관한 통계는 없었습니다. 위의 데이터에서 사고의 대부분은 안전운전 의무 불이행(사방을 살피면서 가지 않음), 보행자 보호의무 위반(횡단보도에서 횡단하는 사람들에 대한 보호), 신호위반에서 대부분의 사고가 발생합니다. 보행자 보호의무 위반이라는 것에 대해서 의아해서 살펴봤더니 흔히들 운전자들은 우회전시에 횡단보도에서 초록불에도 사람들이 없으면 지나가곤 합니다. 거의 신호가 끝났고, 아무도 없으니 그냥 가도 되겠지 라는 생각으로 이런 경우는 굉장히 많고 저도 이렇게 합니다. 그러다가 어린이가 갑자기 튀어 나온다거나 키가 작아서 못 보는 경우가 있는데 이런 경우가 보행자 보호의무 위반사고의 대부분을 차지한다고 나와있었습니다. 또한, 안전운전 의무 불이행의 경우 뉴스기사를 발췌했습니다. “전문가들은 “어린이는 키가 작아 빈번히 운전자의 사각지대에 놓이고, 충동이나 몰입 성향이 강해 갑자기 도로로 뛰어들거나 무단 횡단할 확률이 높아 안전운전 의무 불이행에 따른 교통사고가 많다” **라는 내용입니다. 현재 제가 가져온 데이터는 2018년 데이터이고, 2016년에는 안전운전의무불이행이 어린이교통사고의 62%를 차지하였습니다. 이렇게 우발적으로 사고가 나는 경우가 상당 수였습니다. 결론적으로 “주중 하교시간과 학원 끝나는 시간(16시~20시)에 우발적 사고가 빈번하게 일어났다”라고 정리할 수 있겠습니다. 현재 데이터는 굉장히 단순한 형태로 그래프처럼 시각적인 관찰이 없어도 직관적으로 한 눈에 파악하기 좋은 형태였고 이런 사고의 발생유형으로 볼 때 사고 예방관련서비스를 개발한다면 같은 운전자 입장으로써 다른 운전자들에게 많은 도움이 될 것 같아서 선택하였고, 개발한다면, 네비게이션과 연동하여 “현재 스쿨 존이며 사고 다발 시간대이고, 이러이러한 위험이 있으니 주의해주세요”라고 안내를 한다면, 어떠한 사람이라도 더 주의하기 마련입니다. 때문에 사고예방율을 높인다는 것이 제 생각이었습니다.**

**2번째 데이터: 코로나19 데이터(kaggle)**

<https://www.kaggle.com/kimjihoo/coronavirusdataset#Time.csv>

흔한 주제이지만, 아무래도 요즘 코로나 바이러스로 인해 kaggle사이트에서도 제일 상위권에 랭크 되어있고, 이 데이터를 이용하여 감소추세를 본다면 대략적인 종식시기를 알 수 있지 않을까 싶어서 선택하였습니다. 이렇게 대략적인 종식시기를 알 수 있다면, 사람들은 더 대비를 하고 일상으로 돌아갈 수도 있고, 자영업자의 사업계획에 영향을 준다던지 등등 여러 사회적 효과가 있을 것이라고 생각합니다. 때문에 코로나19라는 흔한 주제이지만 선택하게 되었습니다   
**보고서의 양이 제한돼서 코드는 쥬피터 노트북에 첨부하겠습니다.**

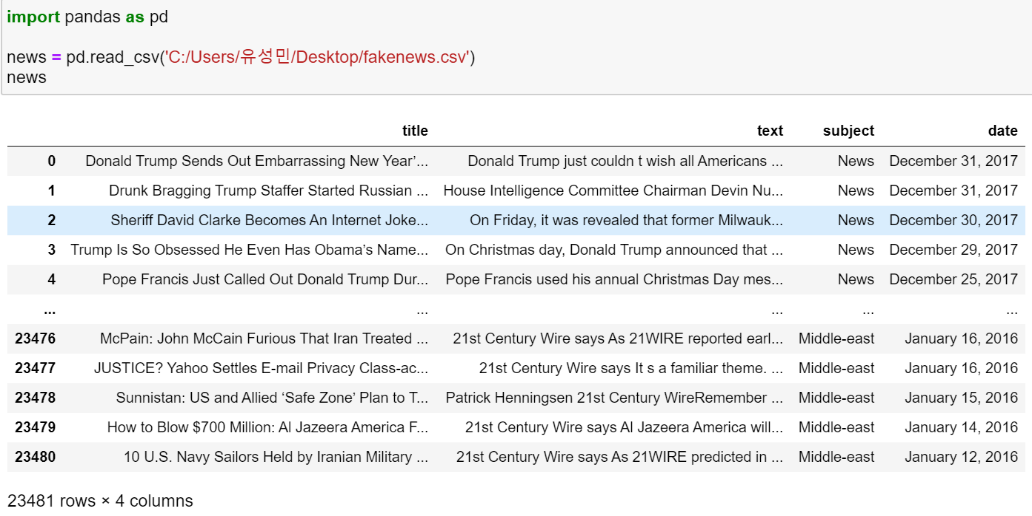
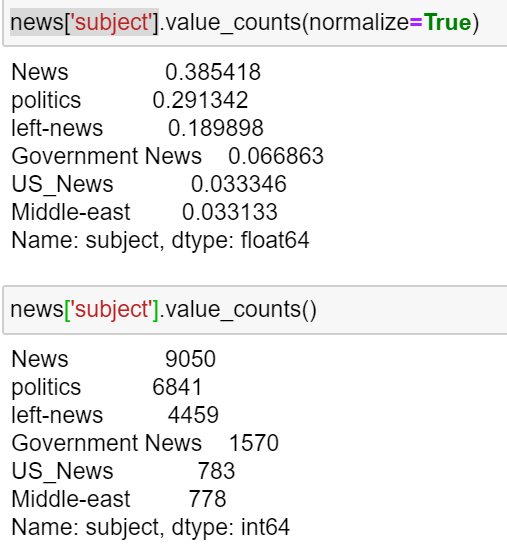
**검사 일자, 데이터 집계 시각 (0 = AM 12:00 / 16 = PM 04:00), 검사 수, 음성 반응, 확진, 완치, 사망 수 입니다. 월이 지남에 따라 얼마나 변하는지 관찰하기 위하여 이 데이터를 선택하였습니다. 1월부터 4월까지 조사하였는데, 확실히 1월 2월 보다는 3월 4월이 눈에 띄게 확진 자 수가 급감했고 2월부터는 검사대비 확진 비율이 4월 말까지 거의 비슷한 수준이었습니다. 아무래도 사회적 거리두기, 마스크 보급 등이 2월부터 시작하였는데 그 영향이 있고 아무래도 제가 의학적 지식이 없다 보니 이 추이만을 보고는 언제쯤 종식이 될지는 판단하기 힘들었습니다. 백신개발이 이뤄지고 있고, 아무래도 여러가지 백신 연구데이터를 수집하여, 백신 임상성공률 같은 것을 본다면 더 좋지 않을까 싶은데, 현재 뚜렷한 백신이 없어서 데이터를 수집하기가 어려웠습니다.**



**2020-1-20 ~ 2020-4-30까지 하루 간격 발생한 데이터입니다. 검사 수, 음성판정, 확진자, 완치자, 사망자 인데, 2월달에 가장 큰 변동을 보였고, 검사 수가 점점 줄어 들고 있는 것이 보입니다. 대부분의 사람들이 가면 음성판정을 받는 것이 보입니다. 확진자 수는 2월에 올라 갔다가 다시 1월달과 비슷한 수치를 보입니다. 완치 판정 받는 사람이 점점 늘어나고 있고, 사망자는 1월부터 거의 비슷한 수준입니다. 당연한 것이겠지만, 이런 추세면 2020년 말 정도에 종식된다고 예측하고, 백신이 나온다면 더 빨리 될 것이고, 신천지 사태같은 큰 사건이 발생한다면 다시 2월-3월 수준이 될 것 같습니다.**

**3번째 데이터: 가짜 뉴스 데이터(kaggle)**

<https://www.kaggle.com/clmentbisaillon/fake-and-real-news-dataset#True.csv>

**요즘 이슈가 되고 있는 가짜 뉴스 데이터입니다. 이 데이터는 가짜 뉴스를 판별하기 위해서 가짜 뉴스로 판명 난 데이터와 진짜 뉴스 데이터를 학습시키고, 새로운 뉴스가 나왔을 때 이 뉴스가 진짜 뉴스인지 가짜 뉴스인지 판별하는 알고리즘을 만드는 데 사용 될 수 있습니다. 다음은 데이터의 일부입니다.**

**해당 데이터를 학습시키기 위해서는 수업시간에는 다뤄지지 않고 소개만 되었던 word embedding이나, transformer등의 언어에 특화된 deep learning 기법이나, 아직 배우지는 못했으나 자연어 처리를 이용하여서 해당 텍스트를 벡터화 하고 비슷한 내용이나 동일한 내용이 나오면 가짜뉴스로 처리하는 등의 작업에 쓰일 수 있을 것입니다. 동일한 사건에 대해서 기자마다 쓰는 내용이 다르겠지만 핵심은 같다는 아이디어에서 출발하면 좋을 것 같습니다. 해당 데이터의 feature에 대해 소개하면, title은 뉴스 제목, text는 뉴스 내용 subject는 뉴스의 주제 토픽 등으로 여러가지가 존재했습니다. politics나 government news 등등 정치나 정책 관련 뉴스가 많은데, 아무래도 정치 관련 기사에서 가짜 뉴스가 많이 나오기 때문에 정치에 관한 기사가 큰 주제로 빠질 만큼 많았습니다.**

**4. 정리**

**3개의 데이터 셋을 보았는데, 이 중에서 교통사고관련 데이터가 방대하여 얻을 정보가 많고, 관련하여 할 수 있는 작업의 영역도 다양하기에 프로젝트 주제로 가장 적절해 보였습니다. 가짜뉴스도 괜찮은 이슈인데, 아직 언어 관련 알고리즘을 배우질 못해서 응용은 어려울 것 같습니다. 마지막으로 코로나19바이러스 관련 데이터인데, 이 데이터는 의학적 배경지식도 부족하고 흔한 주제라서 프로젝트와는 맞지 않다는 결론을 내렸습니다**