기계학습 및 데이터 마이닝 중간고사 대체과제

이름: 김도운

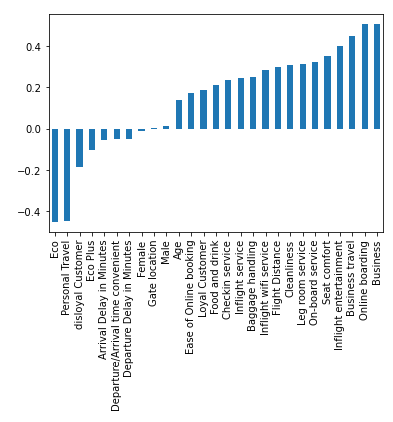
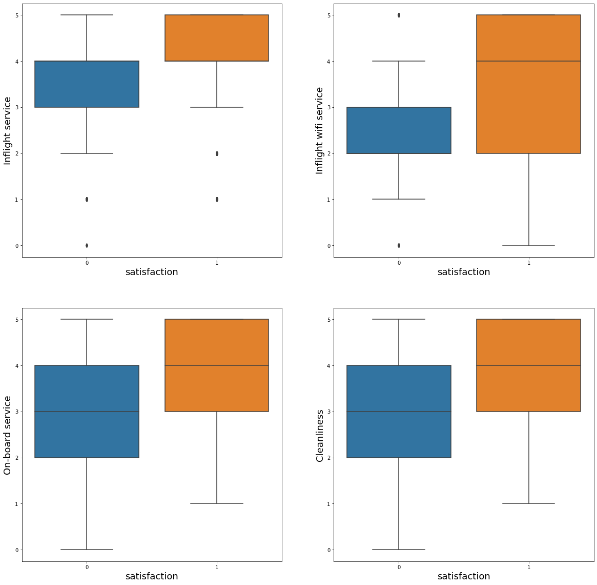
학과: 사이버보안학과

학번: 201520893

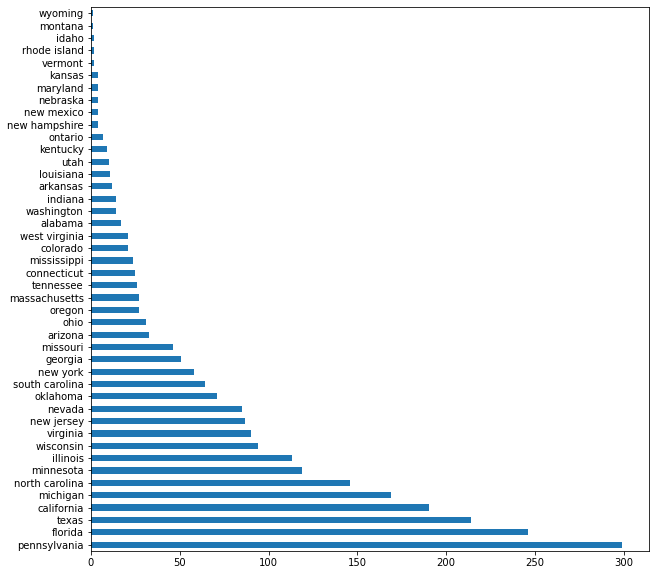
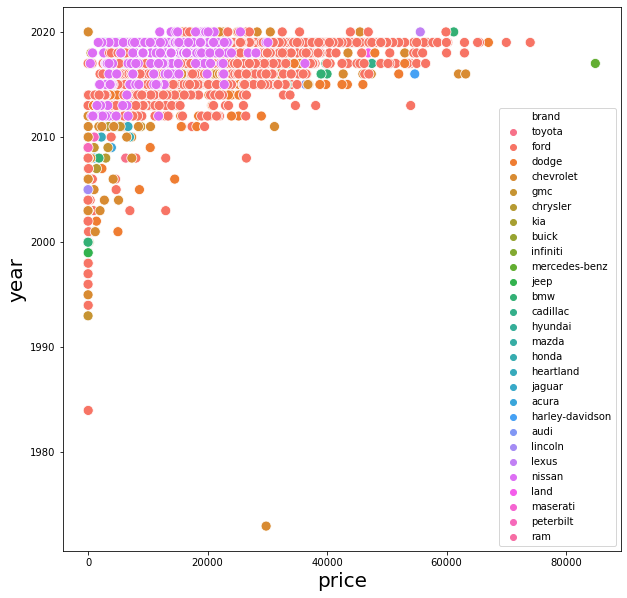
1. **Airline Passenger Satisfaction**

* 이 dataset은 103904명의 서로 다른 승객들에 대해서 성별, 나이, 승객 유형, 여행 유형, 좌석 클래스, 비행 거리, 기내 WIFI 서비스, 출발/도착 시간의 편의성, 온라인 예매의 간편성, Gate 위치, 기내식, 온라인 탑승서비스, 좌석의 편안함, 기내 오락요소(entertainment), 기내 서비스, leg room(다리가 놓아지는 곳) 서비스, 수하물 취급, 체크인 서비스, 청결 정도, 출발 지연시간, 도착 지연시간에 따른 만족도를 조사한 dataset이며 승객 유형, 나이, 여행 유형, 좌석 클래스, 비행 거리, 성별을 제외한 나머지 feature에 대해서는 만족도를 1에서 5로 측정했고(feature값이 0인 경우 해당 feature값이 없음을 의미하며, 숫자가 높을수록 만족도가 높음을 의미한다) 종합적으로 해당 항공편에 대한 만족도를 만족, 중간 또는 불만족의 2개의 class(중간, 불만족은 하나의 클래스)로 분류하였다.
* 총 sample 수는 103904이고, feature수는 22가지이며, 이 중 4가지(성별, 좌석 클래스, 여행 유형, 승객 유형) feature는 category형 변수로, 성별은 남/여, 좌석 클래스는 economy, economy plus, business로 분류되고, 여행 유형은 business travel, personal travel로 분류되며 승객 유형은 loyal customer와 disloyal customer로 분류된다.



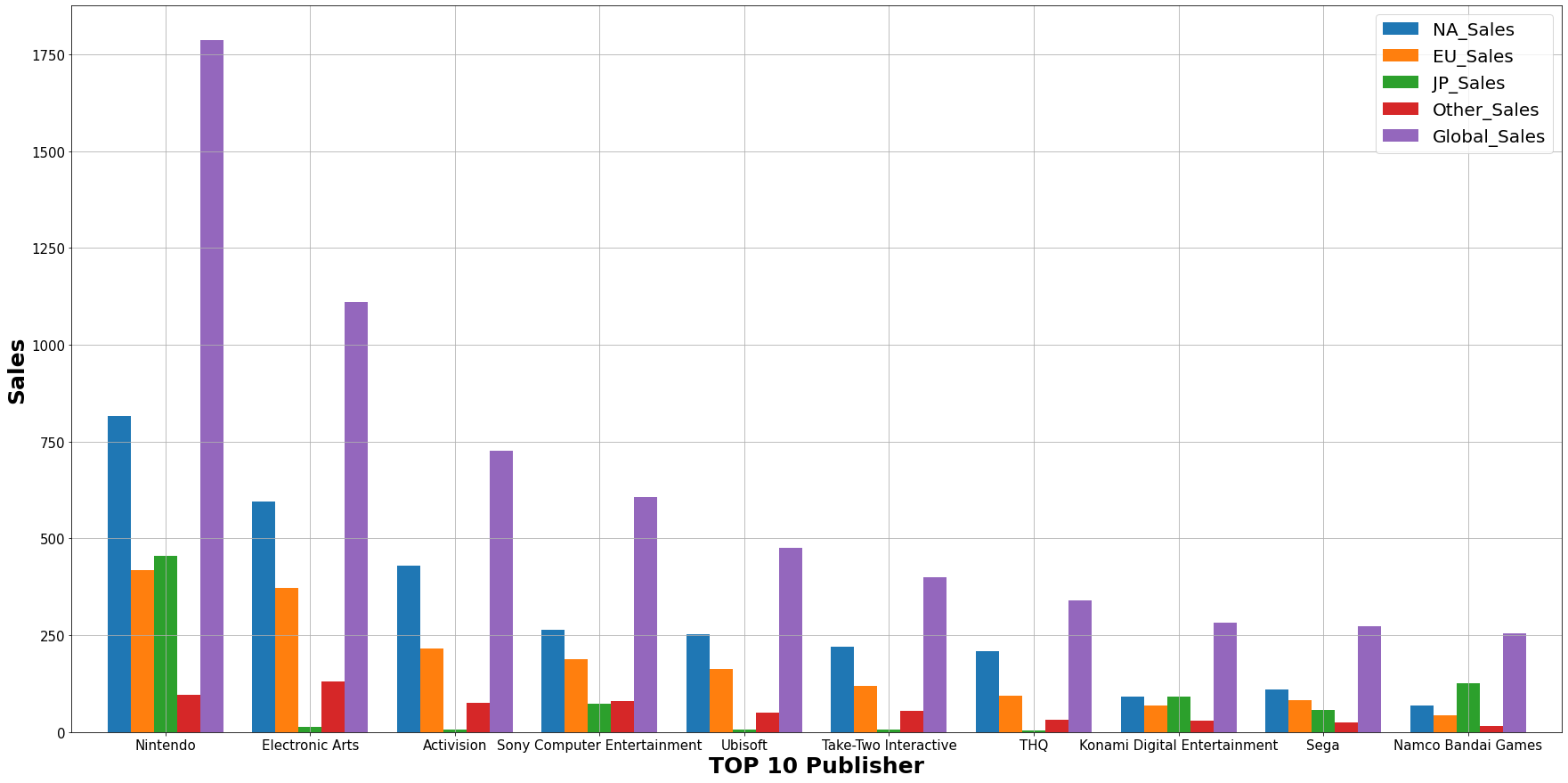
* 위 표는 해당 feature들의 상관관계를 나타내며, 청결도와 높은 상관관계의 feature들은 entertainment, 편안함, 기내식과 같은 서비스 관련 feature들이 많이 연관되어 있음을 알 수 있으며, 그 외에도 제공되는 서비스와 관련된 feature들끼리 상관관계가 높게 측정되었음을 알 수 있다. 또한 온라인예약과 기내 와이파이 서비스의 상관관계가 높은 것을 보아 승객들은 편리할수록, 그리고 제공되는 서비스에 대해 만족도가 높을수록 결과적으로 해당 항공편에 대해서 만족하는 결과를 보여주고 있다. 또한 불편함의 요소가 되는 지연시간이 높아질수록 해당 항공편에 대해 불만족했음을 알 수 있다.
* 위의 bar 그래프는 해당 feature들이 사용자의 만족도에 얼만큼 기여했는지를 보여주고 있다. 성별, 게이트 위치, 지연시간 등은 크게 영향이 없는 것으로 나타나지만, 출발/도착의 지연시간의 경우 지연되지 않는 경우가 훨씬 많으므로(옆의 graph를 보면 지연시간이 없는 경우가 절반 이상이고, 지연시간이 있는 경우, 승객 수가 얼마 없는 것을 확인할 수 있다) 정확하게 측정되었다고 볼 수는 없다. 위에서 말했듯이, 승객의 편안함과 연결되는 feature들(좌석 class, online boarding, WIFI 서비스, 좌석 편안함 정도 등)이 만족도에 크게 영향을 주는 것을 확인할 수 있다. 옆의 boxplot은 ‘청결도’, ‘기내 서비스’, ‘기내 와이파이 서비스’, ‘on boarding’ 서비스에 대해 평가한 점수대별로 만족도와 불만족한 정도를 보여주고 있다. 해당 서비스들에 대해서 4점 이상을 준 승객들은 대부분 만족하였고, 3점 정도로 평가한 승객들은 해당 항공편에 대해 전체적으로 불만족하거나 그저 그랬음을 알 수 있다.
* 이 dataset을 학습시켜 승객들의 항공편에 대한 평가 data를 test하여 승객들이 만족했는지 여부를 알 수 있으며, 전체적으로 낮게 측정된 부분(feature)에 대해 항공사들은 개선할 수 있으며, 승객들의 나이 또는 성별에 따라서 어떤 부분이 만족스러웠는지 통계를 통해 특정 고객을 위한 마케팅 전략을 세워볼 수도 있다.

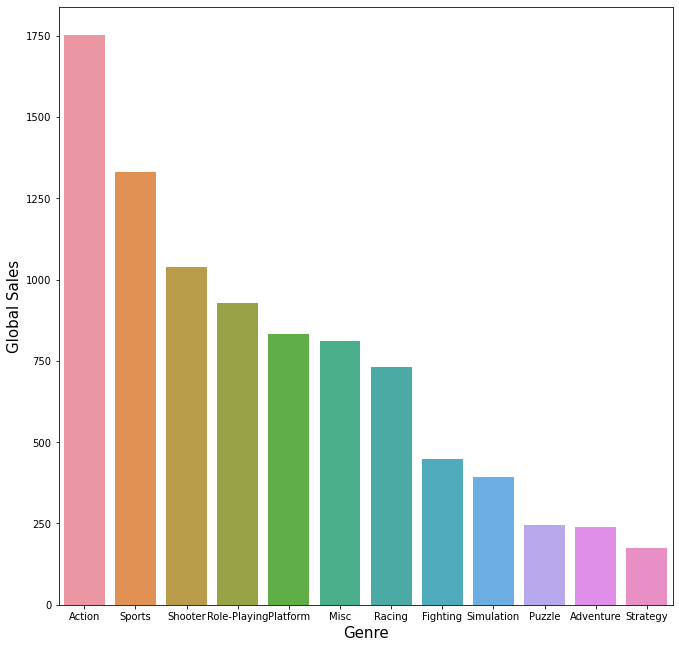
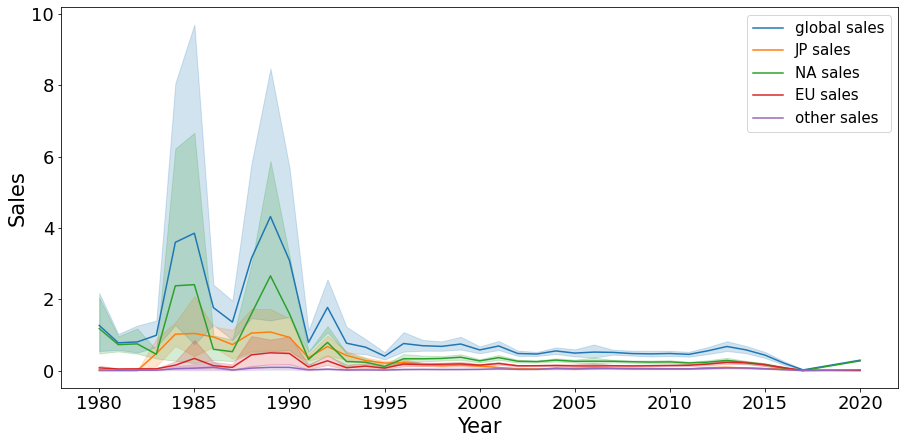
1. **US Cars**

* 이 dataset은 미국에서 판매되고 있는 깨끗한 28개 브랜드의 중고차에 대한 dataset이며, 11개의 feature로 이루어져 있다. 11개의 feature는 가격, 차량 등록 년도, 브랜드, 차량 모델, 차량 색상, 구입 가능한 지역/주, 운행 mile, 차량 등록 번호(vin), 차량 상태(clean or salvage insurance), lot number(차량 식별 번호), 차량 condition(잔여 렌트 기간)으로 이루어져 있다.
* 해당 data를 살펴보기 전에 분석에 필요 없는 feature들(vin, lot, country 등)은 제거한 뒤 describe method를 통해 전체적인 numeric feature들의 분포를 살펴보았다. 오른쪽 표를 보면, 해당 dataset의 year feature의 분포를 통해 차량들이 비교적 최신 차량임을 알 수 있고, 대부분이 최신 차량인 만큼 중고차들의 운행 mile수도 35,000mile로(대략 56,000km) 나쁘지 않다. 가격대 또한 비싼 가격대의 자동차들이 몇 대 있어 보이지만, 평균적으로 20,000$ 근처에 분포하는 것을 알 수 있다.
* 위의 scatter plot은 가격대와 차량 등록 년도에 따라 어떤 브랜드의 차종들이 분포하고 있는지를 나타내고 있다. 위에서 분석했듯이 최근에 등록된 차량들이 대부분 차지하고, 가격대 또한 18,000~20,000$에 많이 분포하고 있음을 확인할 수 있다. 또한 중고차로 등록된 차량들의 브랜드가 주로 ford, Toyota, Lexus, Nissan등인 것으로 주로 일본 차량이 중고차로 많이 사용되는 것을 알 수 있다.
* Barh plot을 살펴보면 중고차가 많이 등록된 state를 알 수 있는데, 펜실베니아, 플로리다, 텍사스, 캘리포니아, 미시간 주 등 미국 동부, 남부, 서부의 큰 주에서 중고차 사업이 많이 진행되고 있음을 알 수 있다.
* 이와 같은 dataset을 사용해 classfication, regression 문제 모두 적용시킬 수 있는데, 예를 들어 분류 문제의 경우 본인이 등록한 차량의 브랜드, 년식에 비해 너무 비싸게 등록된 것은 아닌지 decision tree등을 이용해 다른 dataset들과 비교해 분류해볼 수 있으며, 회귀 문제의 경우 자신이 등록한 중고차의 년식, 운행 마일 수등을 통해 얼마에 등록할지 가격을 예측할 수도 있다.

1. **Video Game Sales**

* 이 dataset은 10만부 이상 판매된 비디오 게임들을 포함하고 있으며, 전체 16598개의 샘플에는 각각 해당 게임의 Ranking, 게임 이름, 플랫폼, 발매 년도, 장르, 출판회사, 북아메리카 판매량, 유럽 판매량, 일본 판매량, 그 외 지역에서의 판매량, 전세계 총 판매량 정보를 feature로 나열하고 있다.



* 위 그래프는 비디오 게임 출판 회사들의 각 지역별 판매량과 판매량이 많은 상위 10개의 회사를 나타내고 있다. 한때 열풍이었던 Nintendo사가 판매량이 가장 높은 것을 확인할 수 있으며, FIFA라는 유명한 축구게임 회사인 Electronic Arts사가 두번째로 높은 것을 확인할 수 있다. 또한 북아메리카 지역량과 유럽 지역에서의 판매량이 많으며, 일본지역도 다른 지역의 인구수에 비해 적은 인구지만 많은 판매량을 기록하고 있는 것을 확인할 수 있다.
* 다음으로 옆의 그래프는 어떤 장르가 판매량이 많았는지를 보여주고 있는데 주로 스포츠, 슈팅 등의 액션 게임들의 판매량이 가장 많았으며, 전략, 모험, 퍼즐 등의 게임들은 비교적 판매량이 적은 것을 확인할 수 있다. 이를 통해 소비자들이 생각을 많이 요구하는 게임이나 시간이 오래 걸리는 게임보다는 좀더 빠르고 역동적인 게임들을 선호한다는 것을 추측할 수 있다.
* 왼쪽의 그래프를 통해 비디오 게임은 80년대 중반부터 90년대 초반까지는 흥행했지만 점점 인터넷이 급속도로 발전하고 모바일 시장이 발전함에 따라서 비디오 게임 시장이 현재는 거의 사라졌음을 확인할 수 있다.
* 이 dataset으로 비디오 게임의 판매량을 예측할 수 있으며, 출시한 게임의 장르에 따라 흥행할 수 있을지 예측이 가능하다. 다만 아쉬운 점은 dataset의 최신 년도에 대한 샘플들이 적기 때문에 과거 년도의 샘플들을 사용하여야 하는데, 이는 현재 시대의 감성과 다를 수도 있기 때문에 정확하게 예측하기를 기대하기는 어려울 수도 있다.

**Dataset 출처 - kaggle**

* Airline Passenger Satisfaction
* <https://www.kaggle.com/teejmahal20/airline-passenger-satisfaction>
* US Cars dataset
* <https://www.kaggle.com/doaaalsenani/usa-cers-dataset>
* Video Games Sales
* <https://www.kaggle.com/gregorut/videogamesales>