**Data Science Competition 2018**

**홍슬기, 장현수**

!!!!! 이건 보고서? 쓴 게 아니라, 데이터 읽어오고 봤던 과정에 대해 서술하고있음 !!!!!

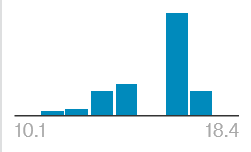
* Dataset : Laptop Prices <https://www.kaggle.com/ionaskel/laptop-prices>
* 그리고 이건 비슷해서 올린건데, 만약 우리가 위에 정한거 하다가 막히면 갈아탈 용도로 알아둠.  
  <https://www.kaggle.com/iliassekkaf/computerparts>

**목표 : 주어진 데이터를 통해 컴퓨터의 가격을 예측하자! 이 컴퓨터의 가성비는 좋을까?**

* Dataset Load and Check : 13 columns and 1303 records  
  : pandas를 이용해 DataFrame형태로 읽어왔다.



아래 설명에서 내가 자꾸 우선순위를 둬서 하나의 column을 보려고 하는 이유는, 가격 책정 때문이다. Class가 너무 많으면 단순하게 158번째 class는 가격을 5.68%를 높여줘! 라는 식의 분석은 너무 세세하고 부정확할 수 있고, 경향을 읽기 힘들다고 생각해서 그러함.  
(물론 그냥 내 생각이얌…나도 이게 데이따분석이 처음이얌..)

* **Columns Info**  
  : 여기서 간략하게 몇%라고 나온거, **나중에 모든 항목에 대해서 제대로 통계내서 도표로 정리할 것.**
  + Unnamed : index (useless)
  + **Company** : Apple, HP, Asus, Dell(23%), Lenovo(23%)… 총 19가지
  + **Product** : 618가지의 제품. MacBook Pro, MacBook Air  
    이건 단순 string으로만 생각할 게 아니라, **첫 띄어쓰기 혹은 -표시 이전의 단어**가 ‘MacBook’, ‘Inspiron’ 이런거라서, 각각을 카테고리별로 그룹을 낸 다음에 뒤에 나오는 글자or숫자로 우선순위를 매긴다는 식의 아이디어도 사용 가능하다.
  + **TypeName** : Notebook(56%), Gaming(16%), 그 외 4가지가 29%
  + **Inches :** 인치.  이건 사이트에 나온건데, 정확한건 역시 도표로 직접 그려야함.
  + **ScreenResolution** : 해상도. ‘1366x768’, ‘Full HD 1920x1080’같이 string으로 되어있다. 무려 40가지. 단순히 40가지의 class(분류)로 볼 것이 아니라, (앞에 글자를 버리고) 숫자 크기만 비교한다면 무엇이 더 고해상도인지 우선순위또한 매길 수 있다.
  + **CPU** : 118가지. 이것도 118가지의 class로 볼 게 아니라, i3 i5 i7등으로만 크게 보거나, 조금 더 세부적으로 본다면 7세대인지 8세대인지 이런식으로 class의 개수를 줄여야한다.  
    (class의 개수를 줄인다는 말의 의미는, 분류가 118개나 되면 복잡하니까 이를 그룹화해서 줄이자는 내 의견.)
  + **Ram** : 9가지. 8GB, 4GB가 대다수이고 9가지. 이것도 우선순위를 매길 수 있다.
  + **Memory :** 39가지. 이건 HDD와 SSD의 조합들 때문에 생각보다 종류가 많다. SSD를 중점적으로 고려하는 식의 분석이 필요하다고 생각함.
  + **GPU :** 110가지. 이건 생각을 같이 해봐야겠다. 인텔, 엔비디아, AMD 등이 있고 뒤에 나오는 글에 따라 사양이 천차만별이니까. (뭐 우리가 굳이 사전에 분류 안해놔도, 나중에 통계내서 순위 보면 알 수 있지 않을까싶긴 하다.)
  + **OpSys :** 운영체제 있는지. 9가지밖에 없음!
  + **Weight :** 이건 뒤에 kg만 빼고, 그냥 float값으로 보면 되겠다! 제일 간단한 column이다!
  + **Price\_euros : 가격. 우리의 예측 목적.**

이 모든 특징에 대한 값을 가지는 1303개의 데이터가 있음. 우리는 상식(당연히 메모리가 큰 게 비싸지! 이런거.)을 적용할 게 아니라, 1303개의 데이터만 보고 진짜 그런건지 검증하는거임. 그니까 상식 믿지말고, 데이터만을 기반으로 분석해야함.

Ex) CPU와 가격사이의 관계 분석. inches와 가격사이의 분석. 등등 1대1.

EX) 가격과의 관계말고, 좋은 cpu를 사용할 때 메모리는 평균적으로 얼마를 사용하는가?  
예를들자면 i7계열은 XX기가, i5계열은 XX기가 등등. 이런식으로 column간의 관계도 분석 가능.  
상관관계들을 시각화하여 heatmap으로도 나타낼 수 있음.

등등… 생각할 수 있는 아이디어 무궁무진하게 많으니까, 계산해보는것들 생기면 공유하자!

**<< 우리가 앞으로 해야할 것 (내생각) >>**

* **자잘한 텍스트 처리 코딩** (ex. Product에서 ‘맥북프로’와 ‘맥북에어’를 ‘맥북’이라는 카테고리에 나눈다거나.) : 조금 더 복잡해지기때문에, 아예 처음부터 하지는 말고, 어느정도 분석 후에 하는 것이 나음.
* **Column들간의 관계 분석하기** : 일단 pandas 사용법 찾아보면 될 것 같아! (배열 행렬 다루느라 numpy 같이 쓸거야. 기본적인 파이썬 리스트, 배열 인덱싱방법 알아두면 답답하지 않게 코딩할 수 있을거야) 혹시 감이 오지 않는다면, 유투브같은데서 데이터셋 하나 정해서 분석 뭐뭐 해보는지 보고 따라서 해보면 좋을 듯!
* **분류기(Classifier) 혹은 Prediction Model 만들기** : 그리고나서 선형 회귀, SVM, 랜덤포레스트 등등 이런거 찾아보면 될거야!
* **이 항목들 중에서 하나 끝내고 다른거 해봐야지 하는 것 보다, 2번 먼저 하다가 막히면 3 하는식으로 번갈아가면서 찾아보는 게 좀 덜 지루할거야 !**
* 모든 작업과정, 통계내는 것은 웬만해서 소스코드를 백업해두자! 이전에했던거 나중에 다시 통계를 보려하면 코딩 귀찮아…  
  몇줄 안되는 코드더라도 ㅇㅇ!