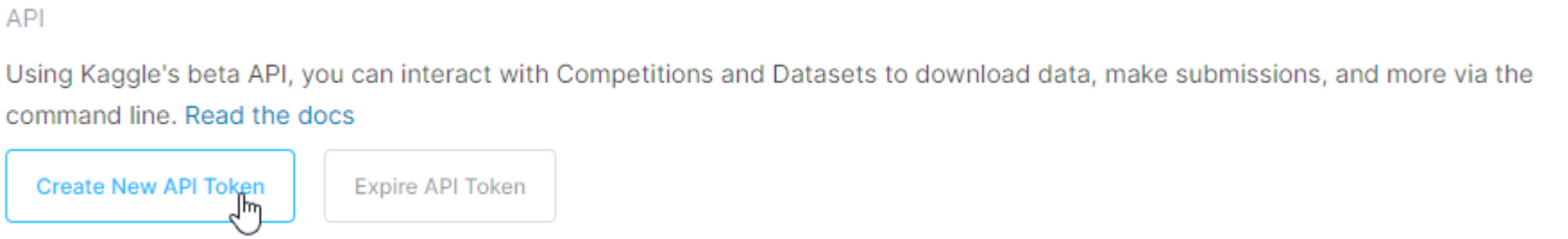
Housing Price 예측 모델 실습

박경규

데이터셋

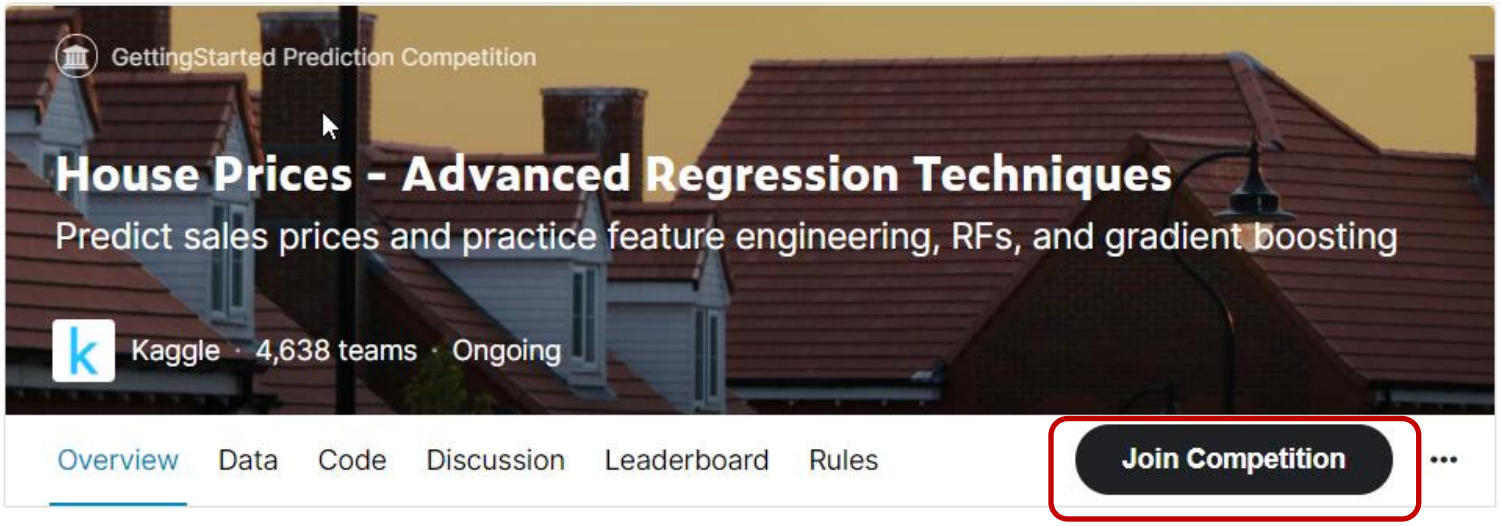
■ kaggle 가입 및 kaggle.json 다운로드

- kaggle 가입 : https://www.kaggle.com/
- kaggle.json 다운로드 : https://www.kaggle.com/<username>/account



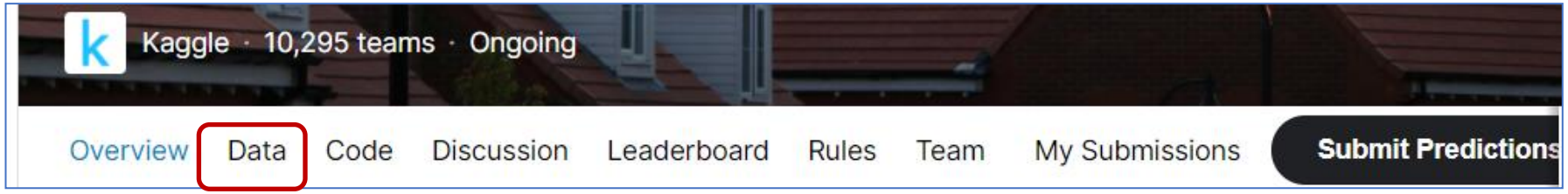
■ 데이터

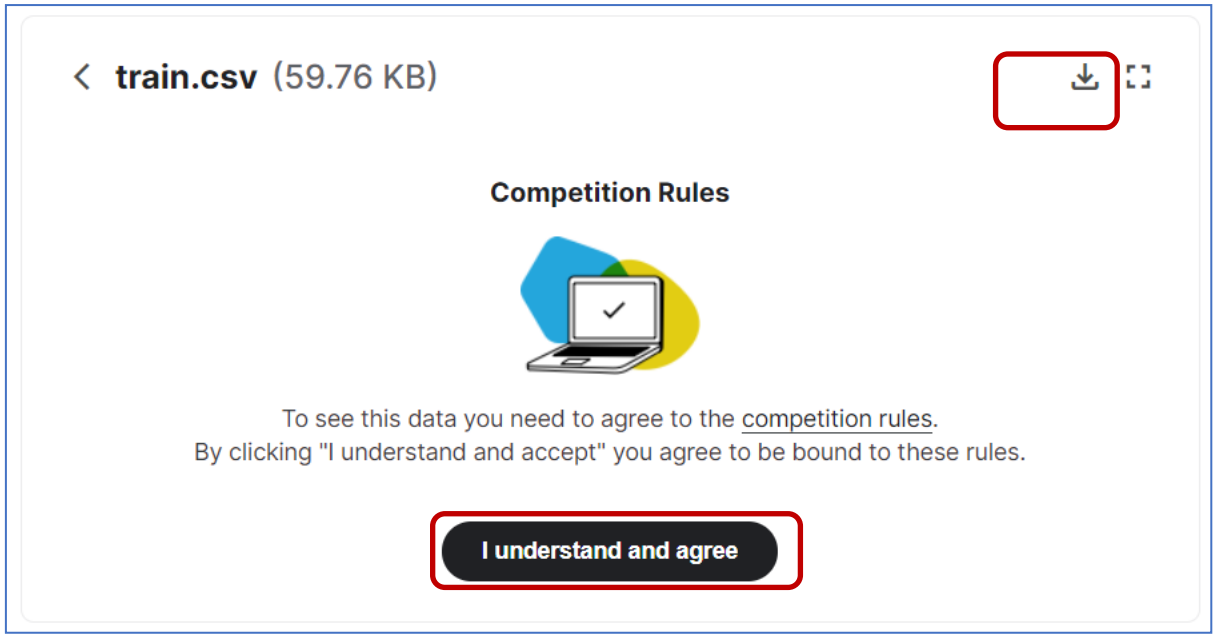
- **House Prices**, https://www.kaggle.com/c/house-prices-advanced-regression-techniques

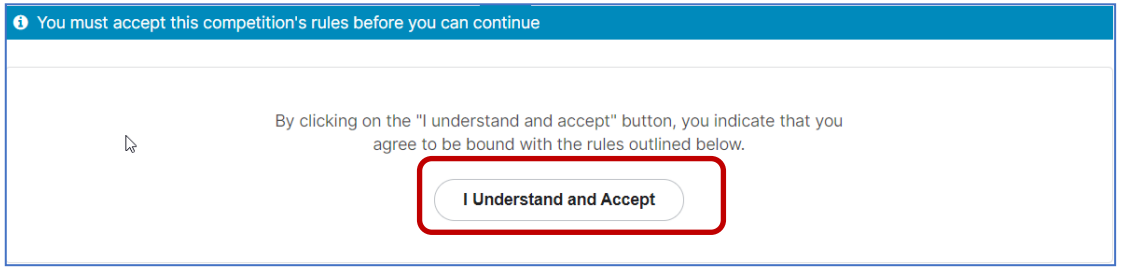


데이터셋

■ kagge 사이트 Data 메뉴 → 다운로드 아이콘 클릭 → agree 버튼 클릭







데이터셋

- MSSubClass: Identifies the type of dwelling involved in the sale.
- MSZoning: Identifies the general zoning classification of the sale.
- LotFrontage: Linear feet of street connected to property
- LotArea: Lot size in square feet
- Street: Type of road access to property
- Alley: Type of alley access to property
- LotShape: General shape of property
- LandContour: Flatness of the property
- Utilities: Type of utilities available
- LotConfig: Lot configuration
- LandSlope: Slope of property
- Neighborhood: Physical locations within Ames city limits
- Condition1: Proximity to various conditions
- Condition2: Proximity to various conditions (if more than one is present)

데이터셋

- BldgType: Type of dwelling
- HouseStyle: Style of dwelling
- OverallQual: Rates the overall material and finish of the house
- OverallCond: Rates the overall condition of the house
- YearBuilt: Original construction date
- YearRemodAdd: Remodel date (same as construction date if no remodeling or additions)
- RoofStyle: Type of roof
- RoofMatl: Roof material
- Exterior1st: Exterior covering on house
- Exterior2nd: Exterior covering on house (if more than one material)
- MasVnrType: Masonry veneer type
- MasVnrArea: Masonry veneer area in square feet
- ExterQual: Evaluates the quality of the material on the exterior
- ExterCond: Evaluates the present condition of the material on the exterior

데이터셋

- Foundation: Type of foundation
- BsmtQual: Evaluates the height of the basement
- BsmtCond: Evaluates the general condition of the basement
- BsmtExposure: Refers to walkout or garden level walls
- BsmtFinType1: Rating of basement finished area
- BsmtFinSF1: Type 1 finished square feet
- BsmtFinType2: Rating of basement finished area (if multiple types)
- BsmtFinSF2: Type 2 finished square feet
- BsmtUnfSF: Unfinished square feet of basement area
- TotalBsmtSF: Total square feet of basement area
- Heating: Type of heating
- HeatingQC: Heating quality and condition
- CentralAir: Central air conditioning
- Electrical: Electrical system

데이터셋

- 1stFlrSF: First Floor square feet
- 2ndFlrSF: Second floor square feet
- LowQualFinSF: Low quality finished square feet (all floors)
- GrLivArea: Above grade (ground) living area square feet
- BsmtFullBath: Basement full bathrooms
- BsmtHalfBath: Basement half bathrooms
- FullBath: Full bathrooms above grade
- HalfBath: Half baths above grade
- Bedroom: Bedrooms above grade (does NOT include basement bedrooms)
- Kitchen: Kitchens above grade
- KitchenQual: Kitchen quality
- TotRmsAbvGrd: Total rooms above grade (does not include bathrooms)
- Functional: Home functionality (Assume typical unless deductions are warranted)

데이터셋

- Fireplaces: Number of fireplaces
- FireplaceQu: Fireplace quality
- GarageType: Garage location
- GarageYrBlt: Year garage was built
- GarageFinish: Interior finish of the garage
- GarageCars: Size of garage in car capacity
- GarageArea: Size of garage in square feet
- GarageQual: Garage quality
- GarageCond: Garage condition
- PavedDrive: Paved driveway
- WoodDeckSF: Wood deck area in square feet
- OpenPorchSF: Open porch area in square feet
- EnclosedPorch: Enclosed porch area in square feet

데이터셋

- 3SsnPorch: Three season porch area in square feet
- ScreenPorch: Screen porch area in square feet
- PoolArea: Pool area in square feet
- PoolQC: Pool quality
- Fence: Fence quality
- MiscFeature: Miscellaneous feature not covered in other categories
- MiscVal: $Value of miscellaneous feature
- MoSold: Month Sold (MM)
- YrSold: Year Sold (YYYY)
- SaleType: Type of sale
- SaleCondition: Condition of sale

참고 - 탐색적 분석과 ML모델링

https://www.kaggle.com/munmun2004/house-prices-for-begginers

I. 데이터 불러오기 및 확인

- a. 데이터 불러오기

II. EDA & FE

- a. Data Processing
- b. concat
- c. 타겟변수 확인
- d. 결측치 확인 및 처리
- e. 순서형과 명목형 데이터 분리
- f. f.파생 변수 생성
- g. 수치형 변수 확인
- h. 범주형 변수 확인
- i. box-cox 변환
- j. 중요 변수 확인

III. 모델링

- a. 단순선형 회귀 & 통계치확인
- b. RobustScaler
- c. KFold
- d. 파라미터 튜닝 & GridSearchCV

IV. 예측

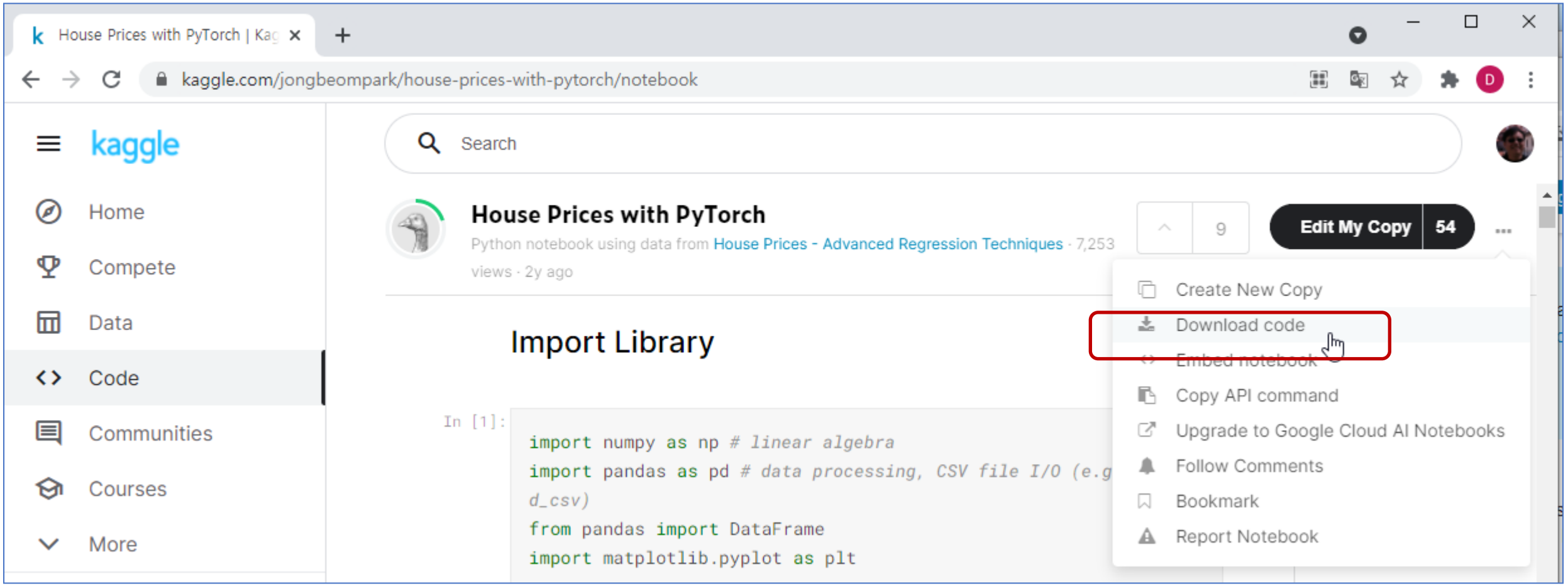
- a. 앙상블
- b. voting
- c. stacking
- c. blending

1. 변수	설명
SalePrice	집값
ID	집을 구분하는 번호
date	집을 구매한 날짜
price	집의 가격
OverallQual	전체 재료 및 마지막 품질
GrLivArea	생활 면적 평방 피트
GarageCars	차고
1stFlrSF	1층 평방 피트
YearBuilt	최초 공사 일
YearRemodAdd	리모델한 날
Fireplaces	벽난로 수
OpenPorchSF	평방 피트 단위의 현관 면적
MasVnrArea	제곱 피트의 벽돌 베니어 면적
LotFrontage	건물에 연결된 거리의 선형 피트
CentralAir	중앙 에어컨
KitchenQual	주방 품질
HeatingQC	난방 품질 및 상태
ExterQual	외부 재료 품질
BsmtQual	지하실의 높이

실습

https://www.kaggle.com/jongbeompark/house-prices-with-pytorch

■ 노트북 파일 다운로드 : Download code



실습

■ 다운로드 받은 노트북파일을 Colab에서 오픈하고 아래 코드를 추가하고 실행

■ Colab에 데이터 저장 디렉토리 생성(실습 목적)

```
import os
DATA_PATH = '../input'

if not os.path.exists(DATA_PATH):
    os.makedirs(DATA_PATH)
```

■ Colab에 kgggle 데이터셋 다운로드

```
! pip install -q kaggle

from google.colab import files
files.upload()
```

```
 ! mkdir ~/.kaggle
 ! cp kaggle.json ~/.kaggle/
 ! chmod 600 ~/.kaggle/kaggle.json

 ! kaggle competitions download -p ../input/ -c house-prices-advanced-regression-techniques
```