Challenge Objective

- Goal: to predict actual transaction price of apartments in Seoul and Busan (Regression Task)
  - The training data consists of 5 years past historical data of apartment transaction prices.
    - e.g., 2018, 2019, 2020, 2021, 2022
  - Predict the transaction prices for the following year, e.g., 2023.

- Any ML models are allowed, not limited to those we learned in our lectures
  - Linear Regression, Random Forest, XGBoost, LightGBM, Deep Learning, and so on.

- Feature engineering and hyperparameter tuning are also key to improve your model.

- Challenge timeline:
  - 4/24 00:00 ~ 5/31 23:59
  - Challenge Link: https://www.kaggle.com/t/b5a7cfbad1844984a803e16a1b9898bd

Dataset

1. train.csv

```
train
```

Target column

	index	apartment_id	city	dong	house_area	built_year	floor	lat	long	transaction_year	transaction_month	transaction_day	PRICE
0	0	0	busan	197	125.865988	1993	5	35.149929	129.006071	2021	7	11~20	229250.80
1	1	0	busan	197	101.647190	1993	12	35.149929	129.006071	2021	10	1~10	215320.00
2	2	0	busan	197	91.511175	1993	6	35.149929	129.006071	2020	3	21~31	161740.00
3	3	0	busan	197	101.647190	1993	13	35.149929	129.006071	2020	5	11~20	199781.80
4	4	0	busan	197	101.647190	1993	4	35.149929	129.006071	2022	6	21~30	219606.40
...	...	...	...	...	...	...	...	...	...	...	...	...	...
329685	329685	4419	seoul	37	101.431912	2014	4	37.452039	127.070842	2022	5	21~31	885070.00
329686	329686	4419	seoul	37	101.431912	2014	14	37.452039	127.070842	2021	10	1~10	826132.00
329687	329687	4419	seoul	37	71.687641	2014	2	37.452039	127.070842	2022	11	21~30	697540.00
329688	329688	4419	seoul	37	137.192013	2014	18	37.452039	127.070842	2020	9	21~30	870656.98
329689	329689	4419	seoul	37	137.192013	2014	18	37.452039	127.070842	2020	8	21~31	871139.20

329690 rows × 13 columns

Dataset

2. test.csv

```
test
```

Target column not available

	index	apartment_id	city	dong	house_area	built_year	floor	lat	long	transaction_year	transaction_month	transaction_day
0	329690	0	busan	197	101.647190	1993	3	35.149929	129.006071	2023	1	21~31
1	329691	0	busan	197	91.511175	1993	12	35.149929	129.006071	2023	2	1~10
2	329692	0	busan	197	125.865988	1993	2	35.149929	129.006071	2023	2	11~20
3	329693	0	busan	197	101.647190	1993	8	35.149929	129.006071	2023	2	21~28
4	329694	0	busan	197	101.647190	1993	13	35.149929	129.006071	2023	3	21~31
...	...	...	...	...	...	...	...	...	...	...	...	...
85092	414782	4419	seoul	37	100.821957	2014	8	37.452039	127.070842	2023	10	21~31
85093	414783	4419	seoul	37	101.431912	2014	11	37.452039	127.070842	2023	10	21~31
85094	414784	4419	seoul	37	121.201627	2014	12	37.452039	127.070842	2023	11	1~10
85095	414785	4419	seoul	37	137.192013	2014	3	37.452039	127.070842	2023	11	21~30
85096	414786	4419	seoul	37	100.821957	2014	4	37.452039	127.070842	2023	11	21~30

85097 rows × 12 columns

Dataset

3. park.csv

```
park
```

	city	gu	dong	park_name	park_type	park_area	park_open_year
0	seoul	23	222	00024j4ovqt3gs6vu6ti	children park	68.992753	2008.0
1	seoul	3	227	00550wcaw8dcg0x4u2f2	neighborhood park	957.700371	1977.0
2	seoul	30	169	01295oe9jrn060obsgfi	neighborhood park	108.990825	2003.0
3	seoul	3	93	02ayq04uy1oz571dp87i	neighborhood park	98.608316	1990.0
4	seoul	21	226	02puz8uy9c7u2x47zhkj	children park	17.146428	1970.0
...	...	...	...	...	...	...	...
1354	seoul	13	194	zudccmxsd2nucbn6e6eo	children park	39.204592	2006.0
1355	seoul	24	84	zv3urzhrt0ghxdn2i6qi	neighborhood park	144.416758	NaN
1356	busan	27	152	zvddij1riig7fxdl2o9k	children park	36.701499	1965.0
1357	seoul	24	84	zxqxz1bym2cqb36ifdiv	children park	68.447060	NaN
1358	seoul	0	10	zyp9kltqez5kpbviuf8l	neighborhood park	105.920725	NaN

1359 rows × 7 columns

Dataset

4. day_care_center.csv

dcc

	city	gu	day_care_name	day_care_type	day_care_baby_num	teacher_num	nursing_room_num	playground_num	CCTV_num	is_commuting_vehicle
0	seoul	2	019dlft759nepobxqft2	national/public	49	9.0	NaN	NaN	NaN	
1	seoul	31	029xz1jtrlpcf9u5sq3i	home	20	4.0	4.0	0.0	0.0	N
2	seoul	3	02e04o0l2br7swfds7jn	private	49	9.0	NaN	NaN	NaN	Y
3	seoul	3	02e04o0l2br7swfds7jn	private	49	9.0	NaN	NaN	NaN	Y
4	busan	14	02e04o0l2br7swfds7jn	home	20	6.0	4.0	0.0	4.0	Y
...	...	...	...	...	...	...	...	...	...	...
7368	seoul	26	zzmuzxowbztizec6dj11	private	28	6.0	NaN	NaN	5.0	N
7369	seoul	3	zzqq55uk81jpkvnrrbni	private	73	11.0	NaN	NaN	NaN	Y
7370	busan	19	zzqq55uk81jpkvnrrbni	home	20	6.0	3.0	0.0	4.0	Y
7371	seoul	3	zzqq55uk81jpkvnrrbni	private	73	11.0	NaN	NaN	NaN	Y
7372	seoul	6	zzqq55uk81jpkvnrrbni	national/public	74	10.0	6.0	1.0	8.0	N

7373 rows × 10 columns

Dataset

5. sample_submission.csv

	index	PRICE
0	329690	0.0
1	329691	0.0
2	329692	0.0
3	329693	0.0
4	329694	0.0
...	...	...
85092	414782	0.0
85093	414783	0.0
85094	414784	0.0
85095	414785	0.0
85096	414786	0.0

85097 rows × 2 columns

6. baseline.csv

	index	PRICE
0	329690	192991.208577
1	329691	166076.068660
2	329692	216900.533457
3	329693	201543.929685
4	329694	200501.292651
...	...	...
85092	414782	820664.339600
85093	414783	772950.635200
85094	414784	797247.557800
85095	414785	803761.439854
85096	414786	760654.025200

85097 rows × 2 columns

Instruction

	index	PRICE
0	329690	0.0
1	329691	0.0
2	329692	0.0
3	329693	0.0
4	329694	0.0
...	...	...
85092	414782	0.0
85093	414783	0.0
85094	414784	0.0
85095	414785	0.0
85096	414786	0.0

85097 rows × 2 columns

- The submission format should look as follows:
- If not, the error will occur in the Kaggle system.

- At most 20 submissions per day are allowed.
  - i.e., 20 * 31 = 620 submissions are maximum.

- We strongly recommend you start early and make sure you have enough time.

- For the final submission, you will be able to make 2 submissions
  - then the system will automatically pick the best score among them.

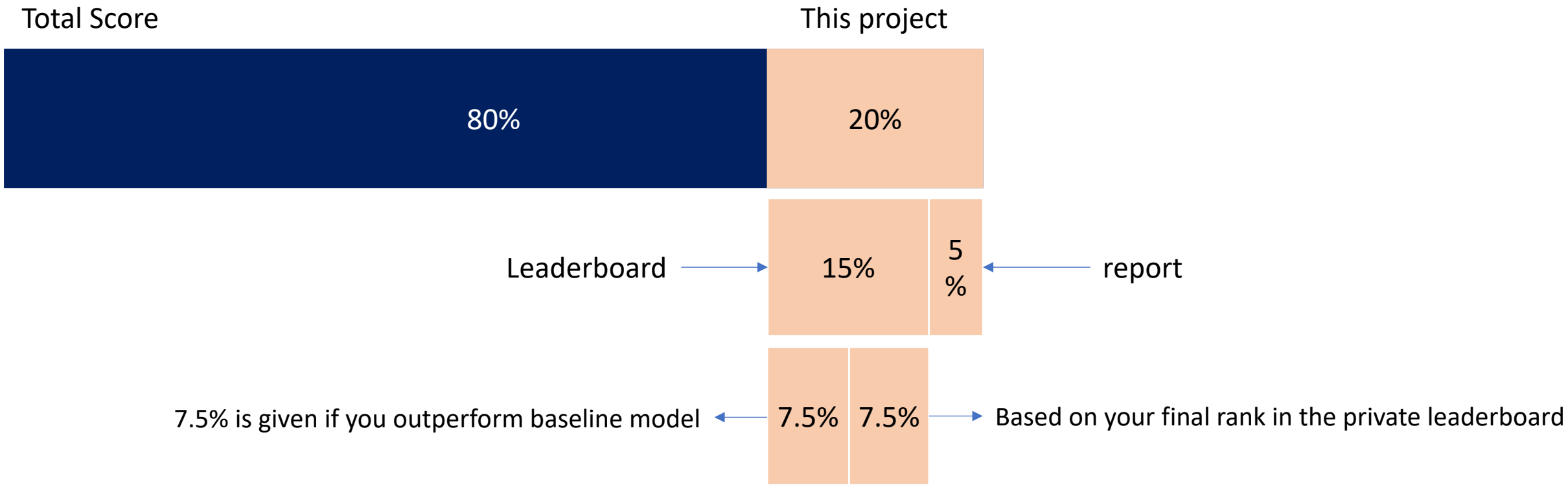
- Video guidelines on how to participate in a Kaggle competition
  - Colab: https://youtu.be/F1P_8qhPVjA
  - Kaggle Notebook: https://youtu.be/sNCwG42r9M4

Evaluation Metric

$$MAE = \frac{\sum_{i=1}^{n} |y_i - \hat{y}_i|}{n}$$

- Mean Absolute Error is calculated as the sum of absolute errors divided by the sample size (n)

Grading



- Note that the leaderboard during the challenge is public leaderboard which is calculated based on 50% of test data.
- The private leaderboard which is calculated based on rest 50 % of test data is your final score

Submission in Kaggle

- Please refer to the video guidelines
  - Colab: https://youtu.be/F1P_8qhPVjA
  - Kaggle Notebook: https://youtu.be/sNCwG42r9M4

Submission in KLMS

- Code
  - The code should be submitted in the provided jupyter notebook format.
    - i.e., code_Gildong Hong_20220000.ipynb
    - You must write and submit your code in the given code format.
  - In particular, the below cell should not be modified. This cell is for loading the input data.

In [3]:
```
## Do not change this cell.
train = pd.read_csv(os.path.join(INPUT_DIR, 'train.csv'))
test = pd.read_csv(os.path.join(INPUT_DIR, 'test.csv'))
park = pd.read_csv(os.path.join(INPUT_DIR, 'park.csv'))
dcc = pd.read_csv(os.path.join(INPUT_DIR, 'day_care_center.csv'))

assert train.shape[0] == 329690 and train.shape[1] == 13, 'Do not change the format of the input dat
a.'
assert test.shape[0] == 85097 and test.shape[1] == 12, 'Do not change the format of the input data.'
assert park.shape[0] == 1359 and park.shape[1] == 7, 'Do not change the format of the input data.'
assert dcc.shape[0] == 7373 and dcc.shape[1] == 10, 'Do not change the format of the input data.'
```

  - The reason for this is to prevent cheating using the original data from DACON.
  - If an assertion error occurs in that cell when Tas evaluate the submitted code, it is considered cheating.
  - We strongly recommend the code to be well documented.

Submission in KLMS

- Report
  - The report should be submitted in a pdf file containing the model explanation and EDA
  - Machine learning model should be built based on data analysis. Apply various methods to analyze the given data
  - Moreover, you will get bonus point if you provide 1) anything interesting from this project, 2) further idea to improve the performance.
  - Unlike the code, there is no specific format for the report.

Rules

- This dataset was released in DACON in 2018, and the winner code is available in here
  - https://dacon.io/competitions/open/21265/codeshare
- You may refer to the code, but any form of plagiarism, including copying code, is strictly prohibited. We will carefully check 1) the similarity between the codes, and 2) correlation between the model predictions.

- Also, this is an individual project, and therefore sharing the code with classmates is strictly prohibited.

- You can use not only the data that are provided **(train.csv, park.csv, daycare.csv)** but also **any other public data** that is available in online. **However, the usage of the original dataset from DACON is not allowed,** which is considered cheating.

Q&A

- Please refer to Appendix of 2023_Project Description.pdf file for the detailed data description.
- If you have any problem or question, feel free to contact us on email or CLASSUM.
- Good Luck !