2017 FAST CAMPUS SCHOOL

DATA SCIENCE SCHOOL PROJECT (1)

/

REGRESSION ANALYSIS

DATA SCIENCE SCHOOL

목 차

1. 팀 공개
2. 데이터셋 소개 (Cross Validation)
3. 주제 설명
4. 주제 선정
5. 타임라인

DATA SCIENCE SCHOOL

팀 공개 A반

1조
박재근, 박재근, 박재근

2조
박재근, 박재근, 박재근

3조
박재근, 박재근, 박재근

4조
박재근, 박재근, 박재근

5조
박재근, 박재근, 박재근

6조
박재근, 박재근, 박재근

7조
박재근, 박재근, 박재근

팀 공개 B반

1조
윤병관, 백승민, 김인수

2조
배광빈, 송세현, 공명구

3조
최규형, 진미나, 염승식

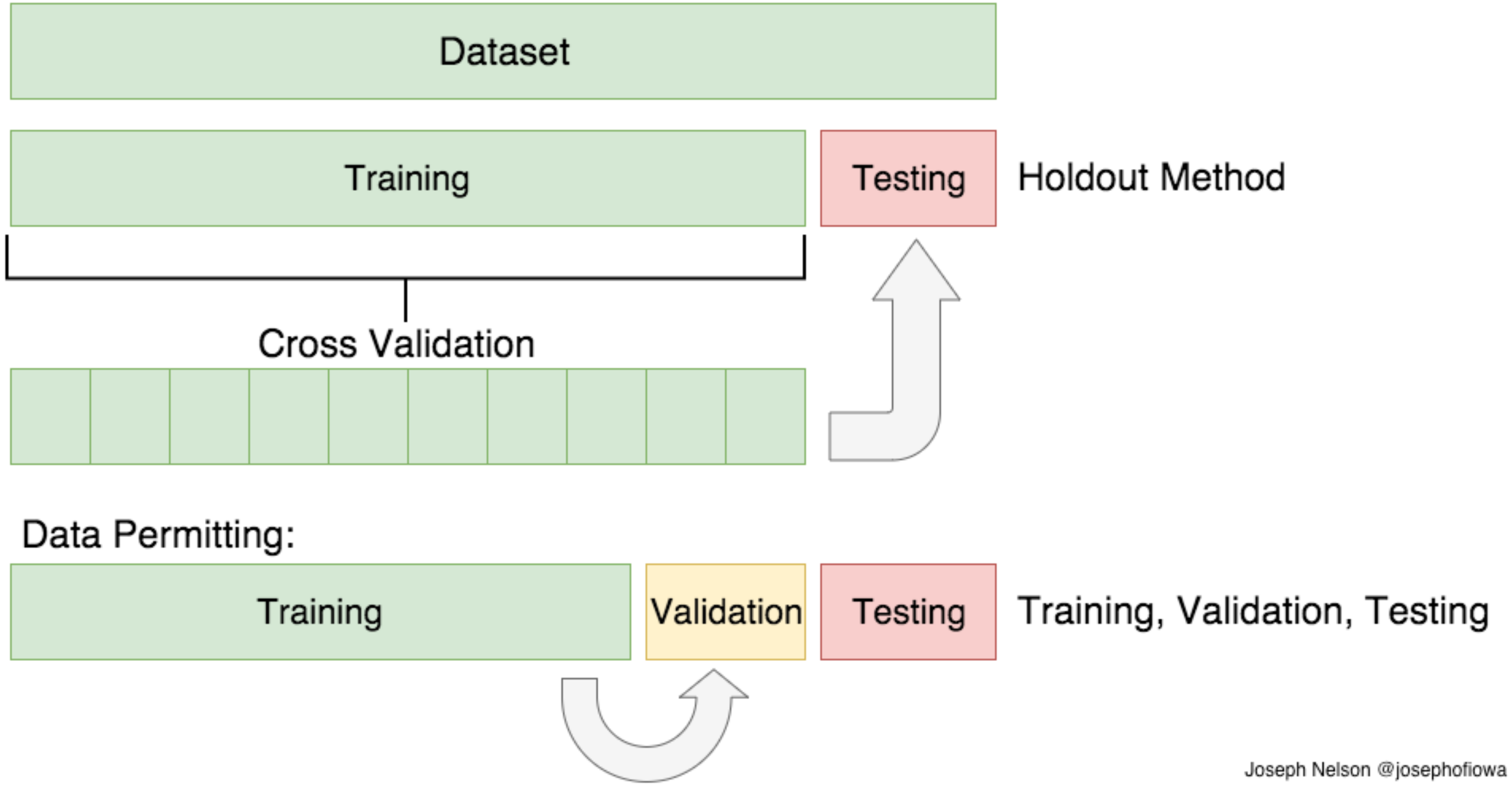
4조
노범용, 유정오, 양영규, 김영법

5조
박상하, 김현규, 한상훈

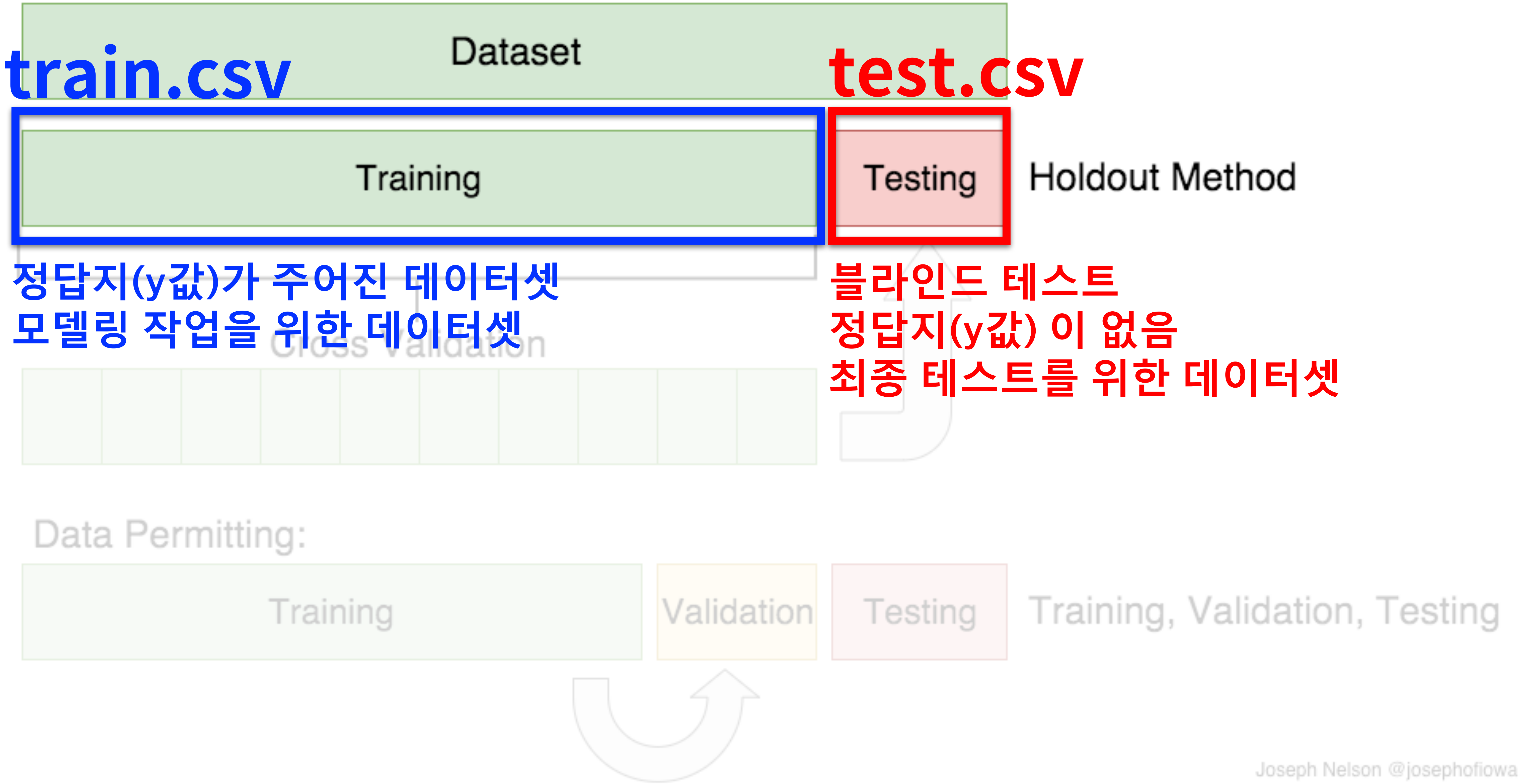
6조
최윤철, 이기훈, 김성희, 안동순

7조
고정욱, 편설인, 이원재

데이터셋 소개


Dataset
Training
Testing
Holdout Method
Cross Validation
Data Permitting:
Training
Validation
Testing
Training, Validation, Testing
Joseph Nelson @josephofiowa

데이터셋 소개 : Train vs Test



train.csv

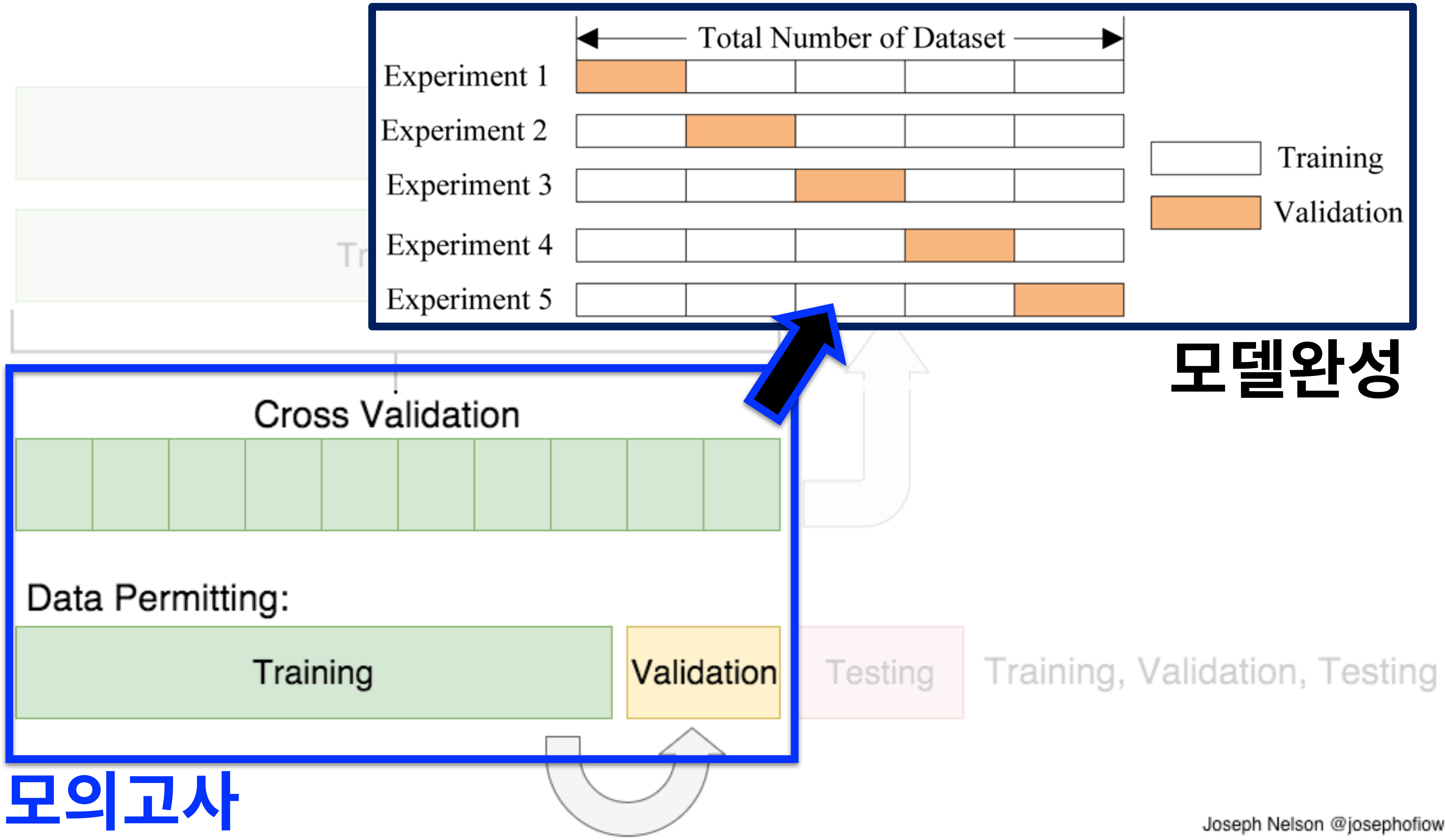
정답지(y값)가 주어진 데이터셋
모델링 작업을 위한 데이터셋

test.csv

블라인드 테스트
정답지(y값) 이 없음
최종 테스트를 위한 데이터셋

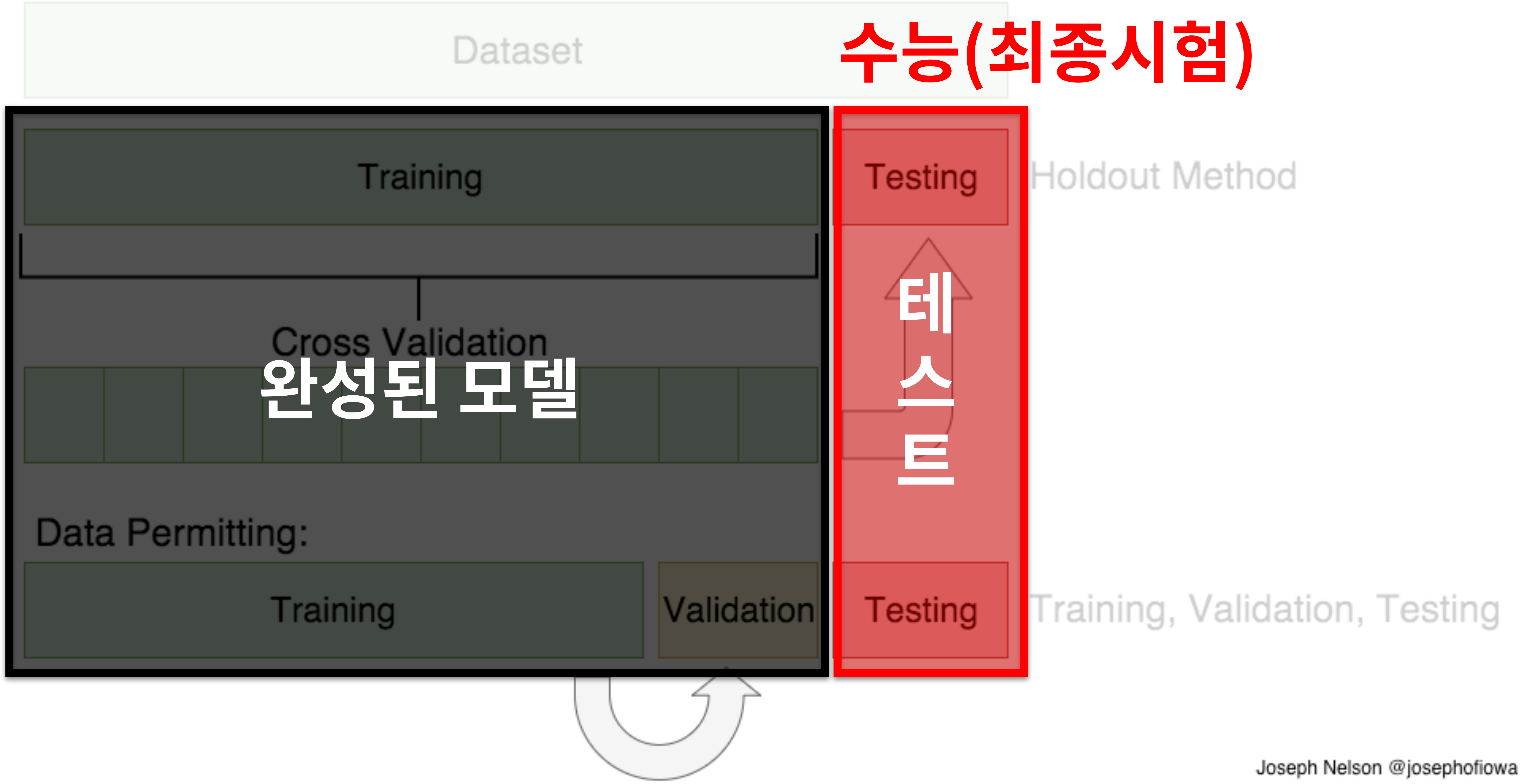
Joseph Nelson @josephofiowa

데이터셋 소개 : Cross Validation



Joseph Nelson @josephofiowa

데이터셋 소개 : Final Test



Joseph Nelson @josephofiowa

주제 설명

1. House Prices: Advanced Regression Techniques

Predict sales prices and practice feature engineering, RFs, and gradient boosting

2. New York City Taxi Trip Duration

Share code and data to improve ride time predictions

3. Sberbank Russian Housing Market

Can you predict realty price fluctuations in Russia’s volatile economy?

4. Toyota Corolla Prices

Predict used Toyota Corolla car prices

House Prices: Advanced Regression Techniques

GOAL

It is your job to predict the sales price for each house.
For each Id in the test set, you must predict the value of the SalePrice variable.

METRIC

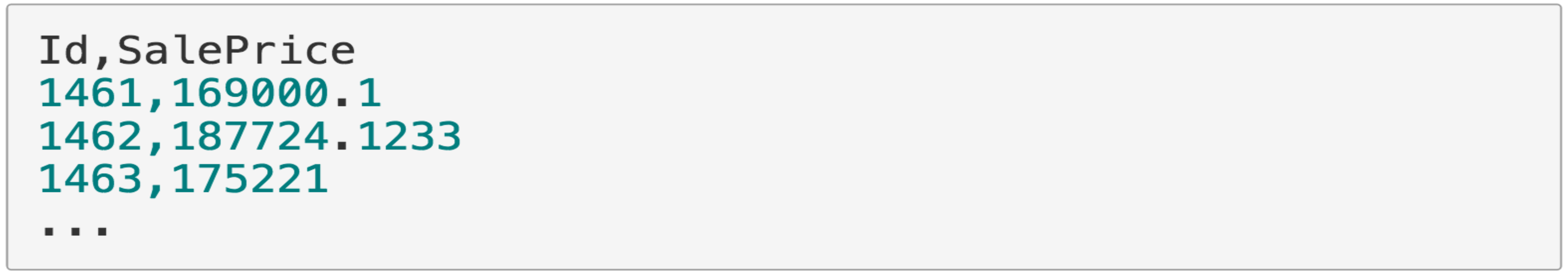
Submissions are evaluated on Root-Mean-Squared-Error (RMSE) between the logarithm of the predicted value and the logarithm of the observed sales price. (Taking logs means that errors in predicting expensive houses and cheap houses will affect the result equally.)

$$\mathrm{RMSD}(\hat{\theta}) = \sqrt{\mathrm{MSE}(\hat{\theta})} = \sqrt{\mathrm{E}((\hat{\theta} - \theta)^2)}.$$

House Prices: Advanced Regression Techniques

SUBMISSION FILE FORMAT

The file should contain a header and have the following format:



```
Id,SalePrice
1461,169000.1
1462,187724.1233
1463,175221
...
```

DATA SIZE

- train.csv : 1030rows, 81colums (@SalePrice)
- test.csv : 430rows, 80colums

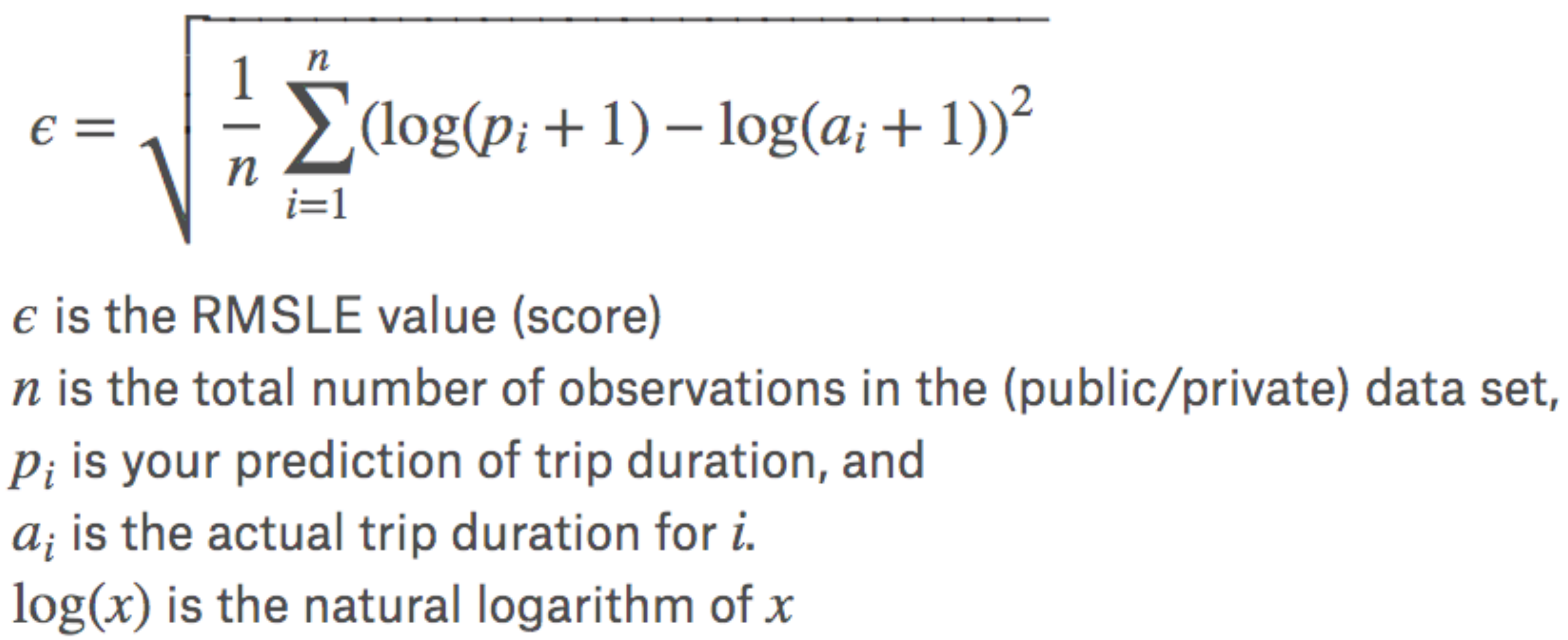
New York City Taxi Trip Duration

GOAL

It is your job to predict the duration time for each id.
For each Id in the test set, you must predict the value of the trip_duration variable.

METRIC

The evaluation metric for this competition is Root Mean Squared Logarithmic Error.
The RMSLE is calculated as



$$\epsilon = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\log(p_i+1)-\log(a_i+1))^2}$$

ϵ is the RMSLE value (score)
n is the total number of observations in the (public/private) data set,
p_i is your prediction of trip duration, and
a_i is the actual trip duration for i.
$\log(x)$ is the natural logarithm of x

New York City Taxi Trip Duration

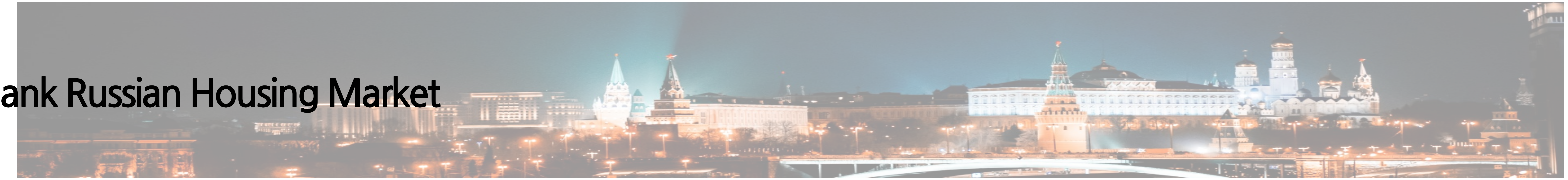
SUBMISSION FILE FORMAT

The file should contain a header and have the following format:

```
id,trip_duration
id00001,978
id00002,978
id00003,978
id00004,978
...
```

DATA SIZE

- train.csv : 701778rows, 11colums (@dropoff_datetime, trip_duration)
- test.csv : 346797rows, 9colums



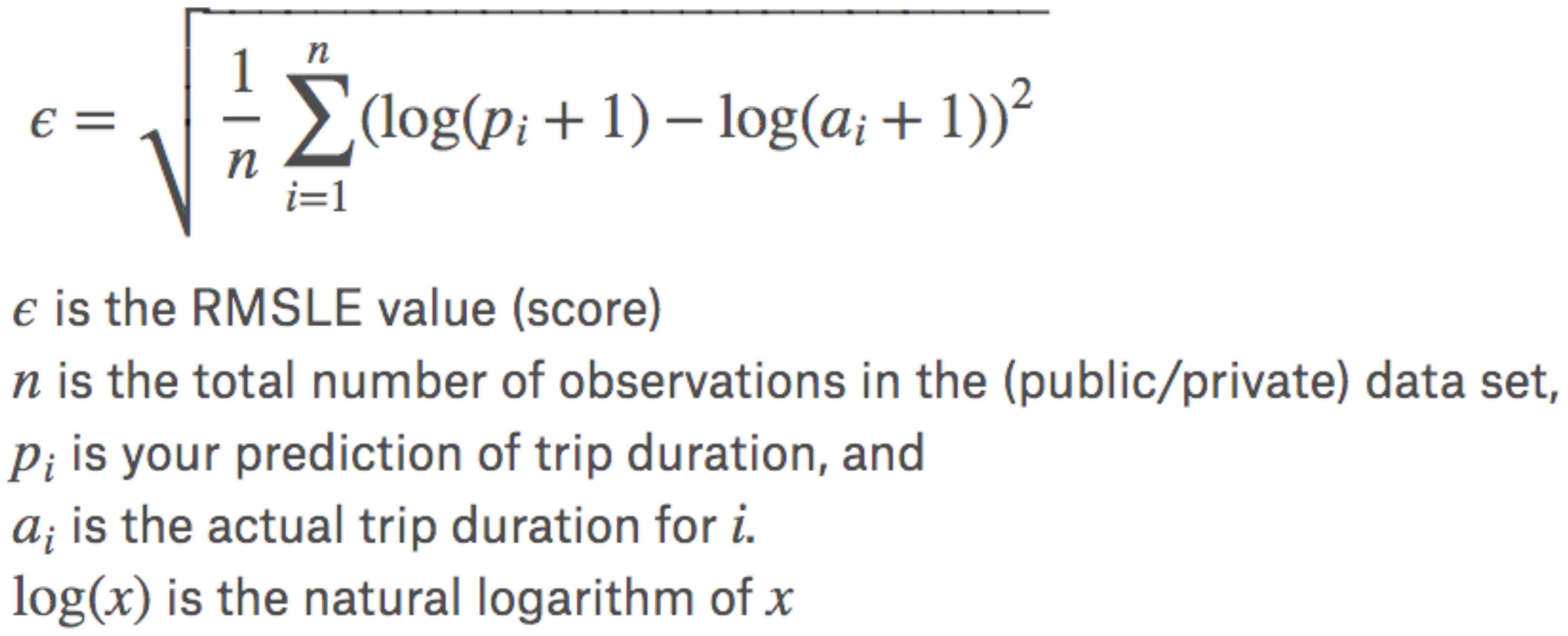
Sberbank Russian Housing Market

GOAL

It is your job to predict the sales price for each house.
For each Id in the test set, you must predict the value of the price_doc variable.

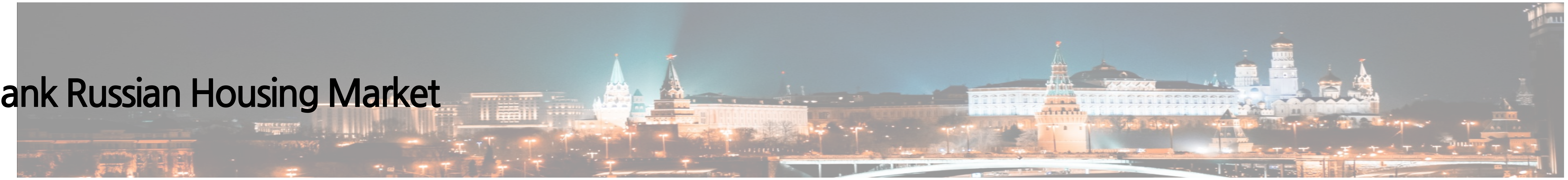
METRIC

The evaluation metric for this competition is Root Mean Squared Logarithmic Error.
The RMSLE is calculated as



$$\epsilon = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (\log(p_i + 1) - \log(a_i + 1))^2}$$

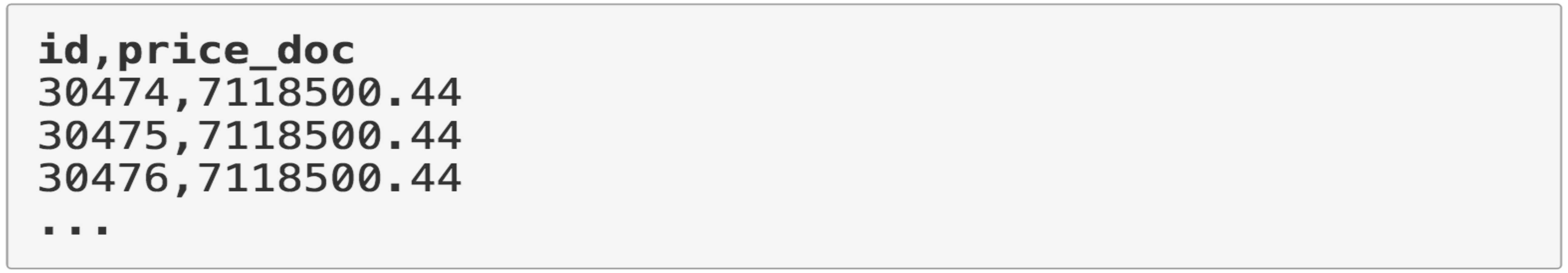
ϵ is the RMSLE value (score)
n is the total number of observations in the (public/private) data set,
p_i is your prediction of trip duration, and
a_i is the actual trip duration for i.
$\log(x)$ is the natural logarithm of x



Sberbank Russian Housing Market

SUBMISSION FILE FORMAT

The file should contain a header and have the following format:



```
id,price_doc
30474,7118500.44
30475,7118500.44
30476,7118500.44
...
```

DATA SIZE

- train.csv : 21570rows, 292colums (@price_doc)
- test.csv : 8901rows, 291colums
- macro.csv : 2484rows, 100colums - data on Russia's macroeconomy and financial sector
  (could be joined to the train and test sets on the "timestamp" column)

Toyota Corolla Prices

GOAL

It is your job to predict the sale price of a used automobile.
For each Id in the test set, you must predict the value of the Price variable.

METRIC

Submissions are evaluated on Root-Mean-Squared-Error (RMSE) between the logarithm of the predicted value and the logarithm of the observed sales price. (Taking logs means that errors in predicting expensive houses and cheap houses will affect the result equally.)

$$\mathrm{RMSD}(\hat{\theta}) = \sqrt{\mathrm{MSE}(\hat{\theta})} = \sqrt{\mathrm{E}((\hat{\theta} - \theta)^2)}.$$

Toyota Corolla Prices

SUBMISSION FILE FORMAT

The file should contain a header and have the following format:

```
Id,Price
3,13950
4,14950
6,12950
7,16900
...
```

DATA SIZE

- train.csv : 1019rows, 39colums (@Price)
- test.csv : 417rows, 38colums

주제 선정

- House Prices
- NYC Taxi
- Sberbank House
- Toyota Corolla

한 주제당 최대 **2조** 까지 선택 가능 **(main)**

한 주제 이상 분석 가능 **(optional)**

데이터 다운로드

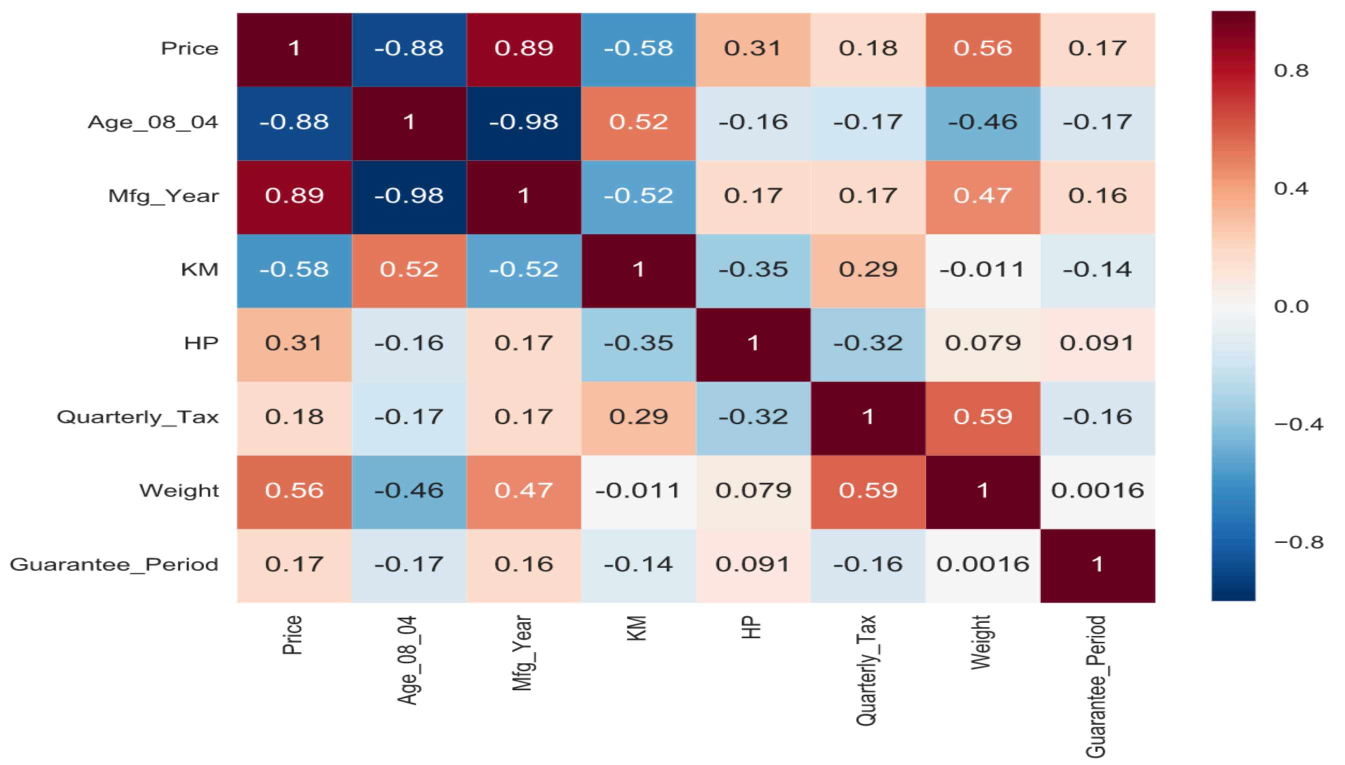
GITHUB | github.com/JKeun/dss-regression-datasets

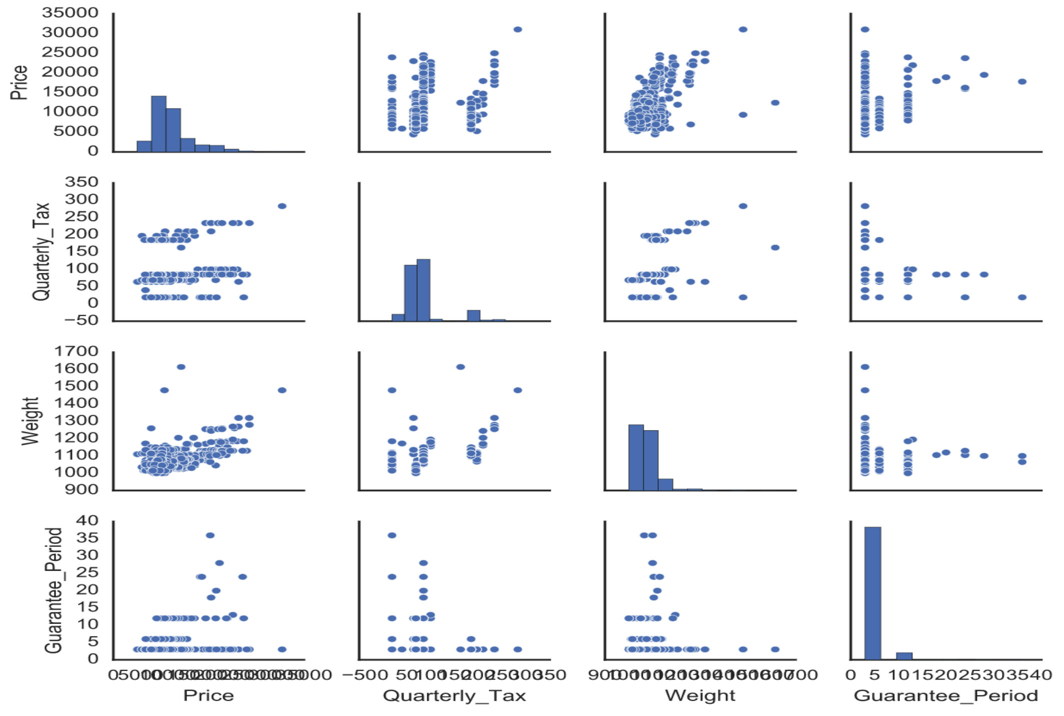
타임라인

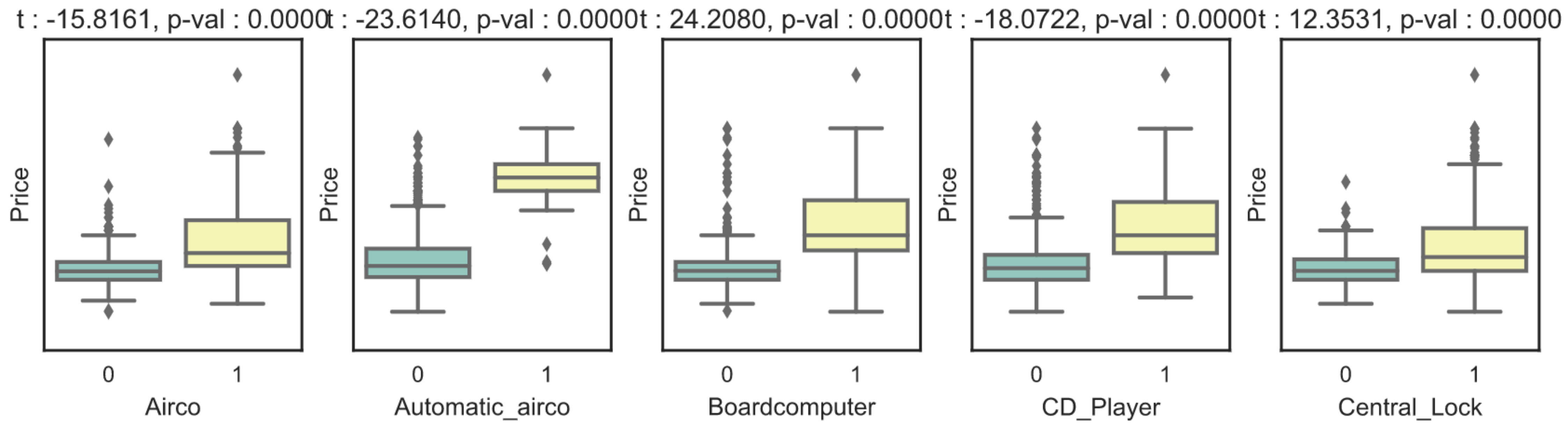
~10/10

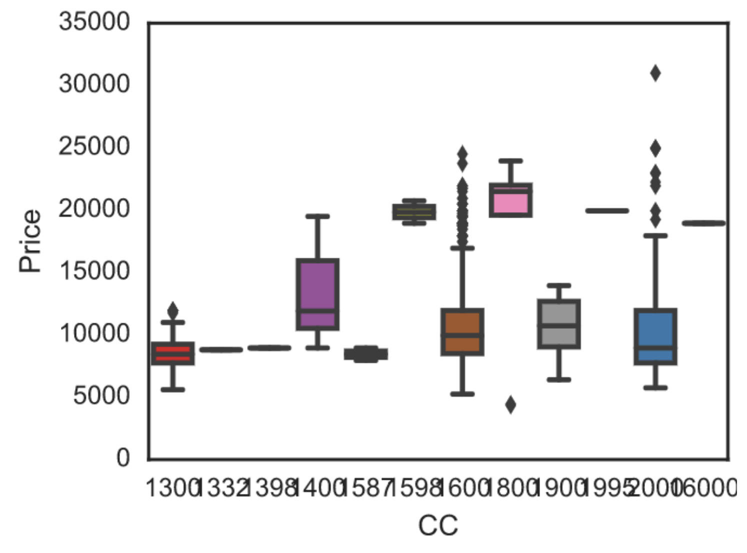
EDA

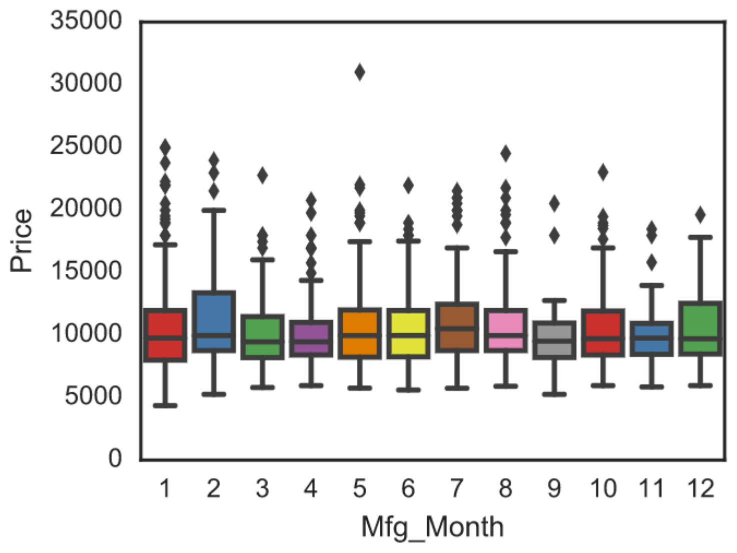
Exploratory data analysis

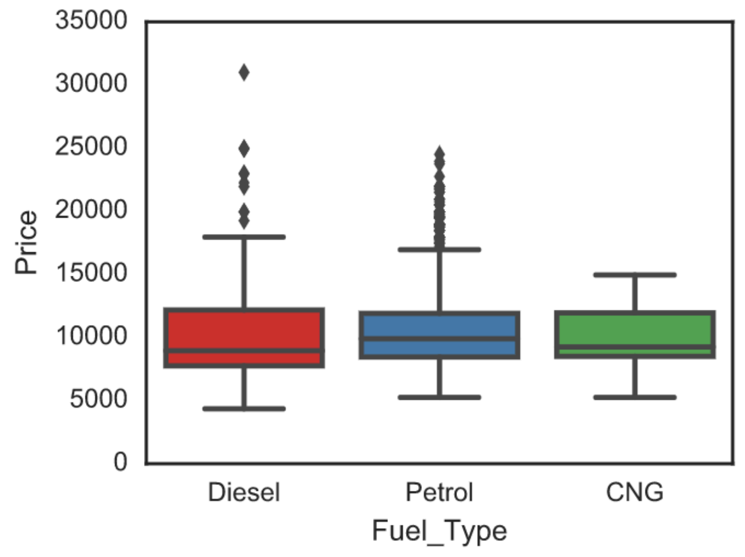








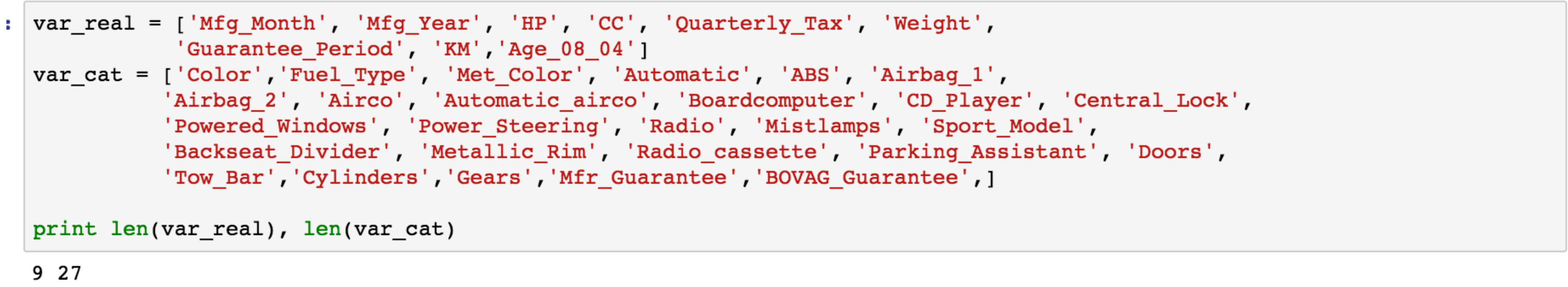




타임라인

~10/14

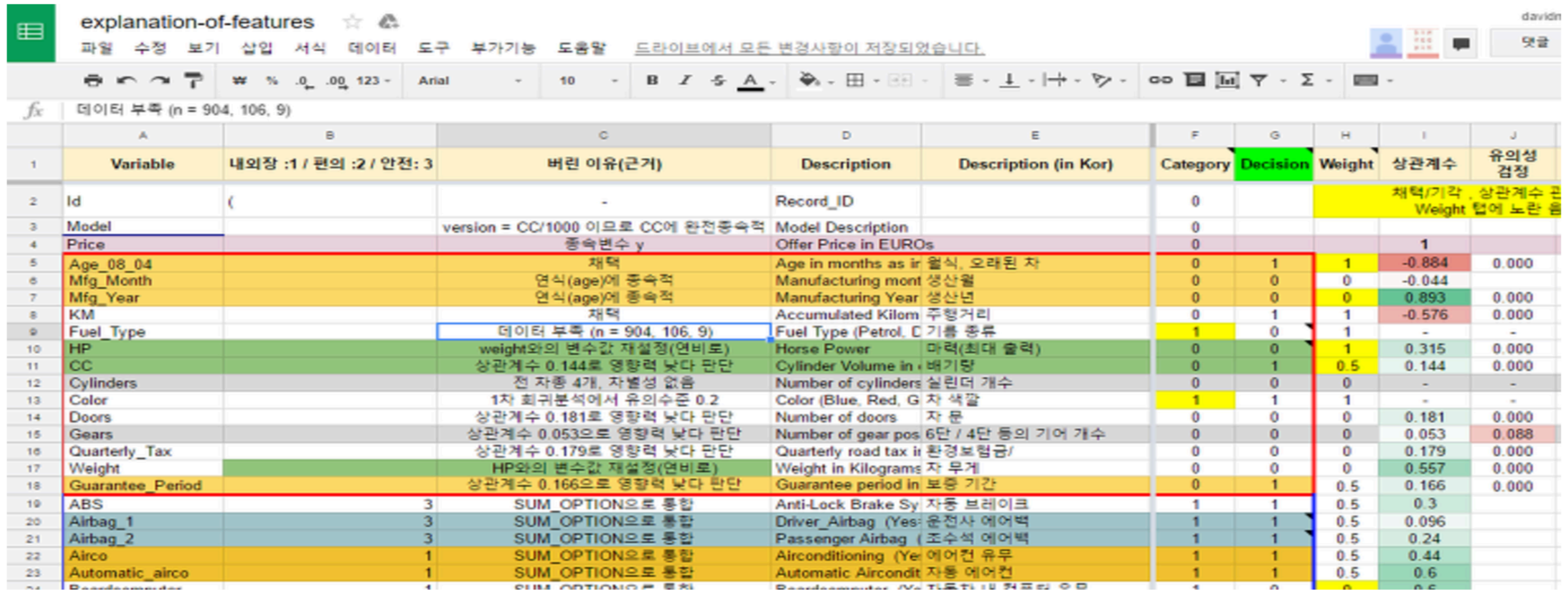
FEATURE SELECTION



```
var_real = ['Mfg_Month', 'Mfg_Year', 'HP', 'CC', 'Quarterly_Tax', 'Weight',
            'Guarantee_Period', 'KM','Age_08_04']
var_cat = ['Color','Fuel_Type', 'Met_Color', 'Automatic', 'ABS', 'Airbag_1',
           'Airbag_2', 'Airco', 'Automatic_airco', 'Boardcomputer', 'CD_Player', 'Central_Lock',
           'Powered_Windows', 'Power_Steering', 'Radio', 'Mistlamps', 'Sport_Model',
           'Backseat_Divider', 'Metallic_Rim', 'Radio_cassette', 'Parking_Assistant', 'Doors',
           'Tow_Bar','Cylinders','Gears','Mfr_Guarantee','BOVAG_Guarantee',]

print len(var_real), len(var_cat)
```

```
9 27
```


explanation-of-features
파일 수정 보기 삽입 서식 데이터 도구 부가기능 도움말
드라이브에서 모든 변경사항이 저장되었습니다.
Arial
10
davidn
댓글

fx 데이터 부족 (n = 904, 106, 9)

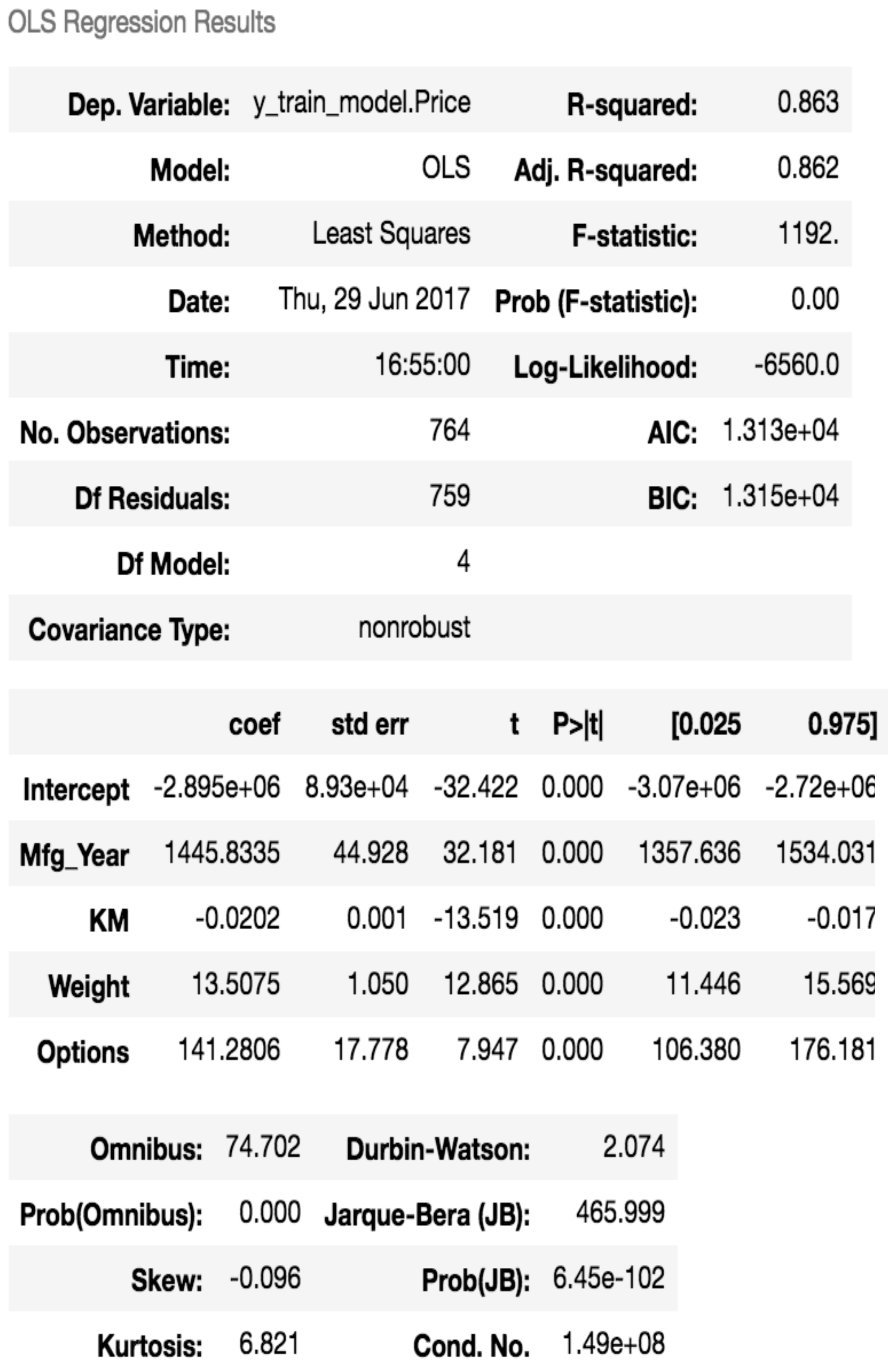
	A	B	C	D	E	F	G	H	I	J
1	Variable	내외장 :1 / 편의 :2 / 안전: 3	버린 이유(근거)	Description	Description (in Kor)	Category	Decision	Weight	상관계수	유의성 검정
2	Id	(	-	Record_ID		0		채택/기각 , 상관계수 Weight 탭에 노란		
3	Model		version = CC/1000 이므로 CC에 완전종속적	Model Description		0				
4	Price		종속변수 y	Offer Price in EUROs		0			1	
5	Age_08_04		채택	Age in months as ir	월식, 오래된 차	0	1	1	-0.884	0.000
6	Mfg_Month		연식(age)에 종속적	Manufacturing mont	생산월	0	0	0	-0.044	
7	Mfg_Year		연식(age)에 종속적	Manufacturing Year	생산년	0	0	0	0.893	0.000
8	KM		채택	Accumulated Kilom	주행거리	0	1	1	-0.576	0.000
9	Fuel_Type		데이터 부족 (n = 904, 106, 9)	Fuel Type (Petrol, D	기름 종류	1	0	1	-	-
10	HP		weight와의 변수값 재설정(연비로)	Horse Power	마력(최대 출력)	0	0	1	0.315	0.000
11	CC		상관계수 0.144로 영향력 낮다 판단	Cylinder Volume in	배기량	0	1	0.5	0.144	0.000
12	Cylinders		전 자종 4개, 자별성 없음	Number of cylinders	실린더 개수	0	0	0	-	-
13	Color		1차 회귀분석에서 유의수준 0.2	Color (Blue, Red, G	차 색깔	1	1	1	-	-
14	Doors		상관계수 0.181로 영향력 낮다 판단	Number of doors	차 문	0	0	0	0.181	0.000
15	Gears		상관계수 0.053으로 영향력 낮다 판단	Number of gear pos	6단 / 4단 등의 기어 개수	0	0	0	0.053	0.088
16	Quarterly_Tax		상관계수 0.179로 영향력 낮다 판단	Quarterly road tax ir	환경보험금/	0	0	0	0.179	0.000
17	Weight		HP와의 변수값 재설정(연비로)	Weight in Kilograms	차 무게	0	0	0	0.557	0.000
18	Guarantee_Period		상관계수 0.166으로 영향력 낮다 판단	Guarantee period in	보증 기간	0	1	0.5	0.166	0.000
19	ABS	3	SUM_OPTION으로 통합	Anti-Lock Brake Sy	자동 브레이크	1	1	0.5	0.3	
20	Airbag_1	3	SUM_OPTION으로 통합	Driver_Airbag (Yes	운전사 에어백	1	1	0.5	0.096	
21	Airbag_2	3	SUM_OPTION으로 통합	Passenger Airbag (	조수석 에어백	1	1	0.5	0.24	
22	Airco	1	SUM_OPTION으로 통합	Airconditioning (Ye	에어컨 유무	1	1	0.5	0.44	
23	Automatic_airco	1	SUM_OPTION으로 통합	Automatic Aircondit	자동 에어컨	1	1	0.5	0.6	

타임라인

~10/16

MODELING

OLS Regression

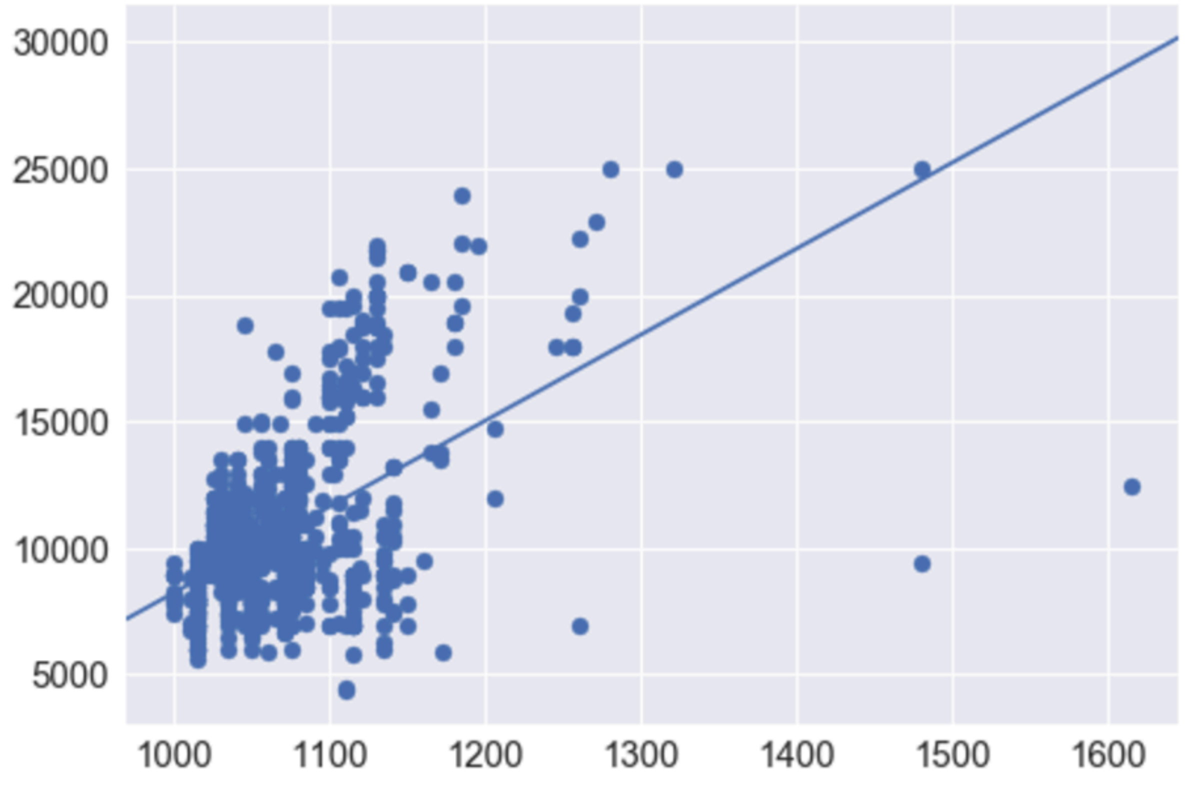


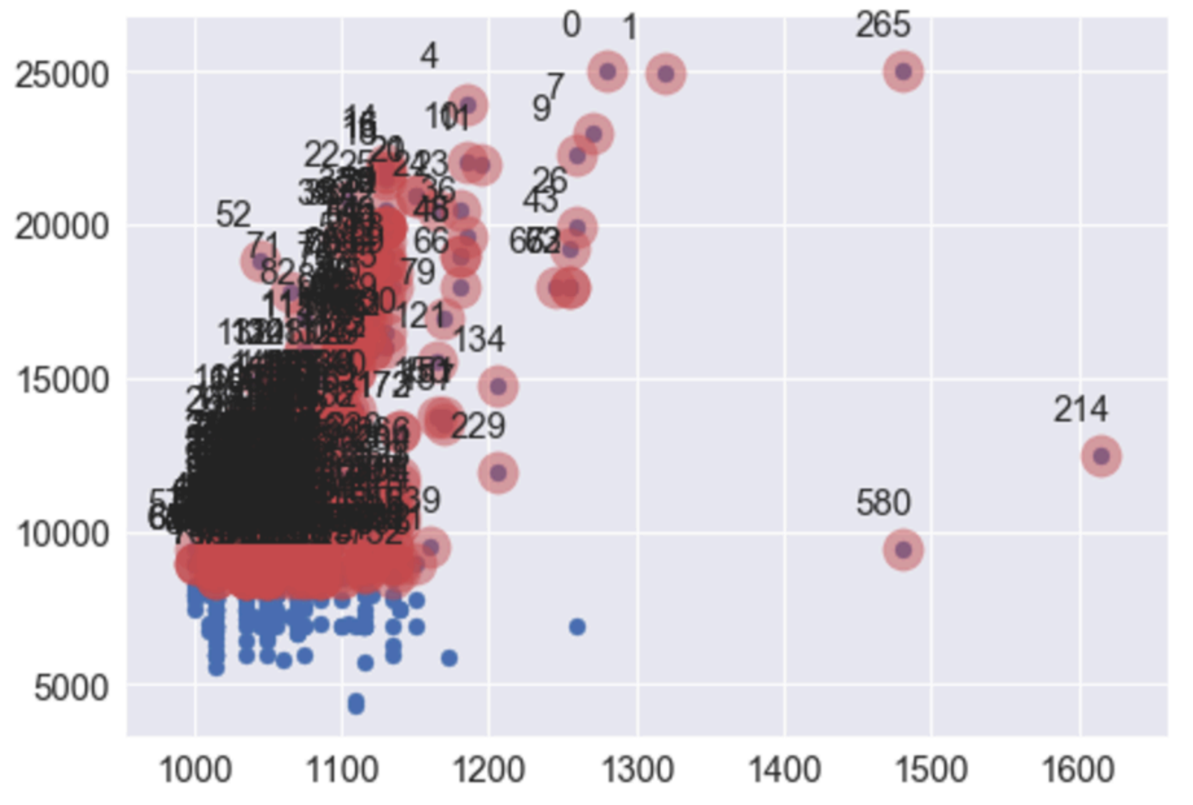
OLS Regression Results

Dep. Variable:	y_train_model.Price	**R-squared:**	0.863
Model:	OLS	**Adj. R-squared:**	0.862
Method:	Least Squares	**F-statistic:**	1192.
Date:	Thu, 29 Jun 2017	**Prob (F-statistic):**	0.00
Time:	16:55:00	**Log-Likelihood:**	-6560.0
No. Observations:	764	**AIC:**	1.313e+04
Df Residuals:	759	**BIC:**	1.315e+04
Df Model:	4		
Covariance Type:	nonrobust		

	coef	**std err**	**t**	**P>|t|**	**[0.025**	**0.975]**
Intercept	-2.895e+06	8.93e+04	-32.422	0.000	-3.07e+06	-2.72e+06
Mfg_Year	1445.8335	44.928	32.181	0.000	1357.636	1534.031
KM	-0.0202	0.001	-13.519	0.000	-0.023	-0.017
Weight	13.5075	1.050	12.865	0.000	11.446	15.569
Options	141.2806	17.778	7.947	0.000	106.380	176.181

Omnibus:	74.702	**Durbin-Watson:**	2.074
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	465.999
Skew:	-0.096	**Prob(JB):**	6.45e-102
Kurtosis:	6.821	**Cond. No.**	1.49e+08





DATA SCIENCE SCHOOL

타임라인

~10/21

MODEL SELECTION

Milestone #12. 최종 회귀모델 결정

Price = 19,230 - 138.583(Age_08_04) - 0.0165(KM) + 546.796(InOut) - 396.891(Safe)

타임라인

~10/25

PROJECT 발표

10/24 까지 프로젝트 **발표자료** 완성 (Jupyter Notebook, PPT)

10월 25일 수요일 프로젝트 발표

END

THANK YOU ; -)