

2017 FAST CAMPUS SCHOOL

# DATA SCIENCE SCHOOL PROJECT ( 1 )

/

## REGRESSION ANALYSIS

박재근 매니저

DATA SCIENCE SCHOOL

# 목 차

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

## 팀 공개 A반

1조

**이신형, 신신호, 강민구**

3조

**강창기, 김경윤, 지성민**

5조

**조성빈, 심호섭, 엄인성**

7조

**조한준, 이정언, 김동현**

2조

**최정혁, 권순호, 박현, 신정원**

4조

**이세영, 이규태, 허성현**

6조

**조현윤, 정하연, 이상협**

## 팀 공개 B반

1조  
**윤병관, 백승민, 김인수**

3조  
**최규형, 진미나, 염승식**

5조  
**박상하, 김현규, 한상훈**

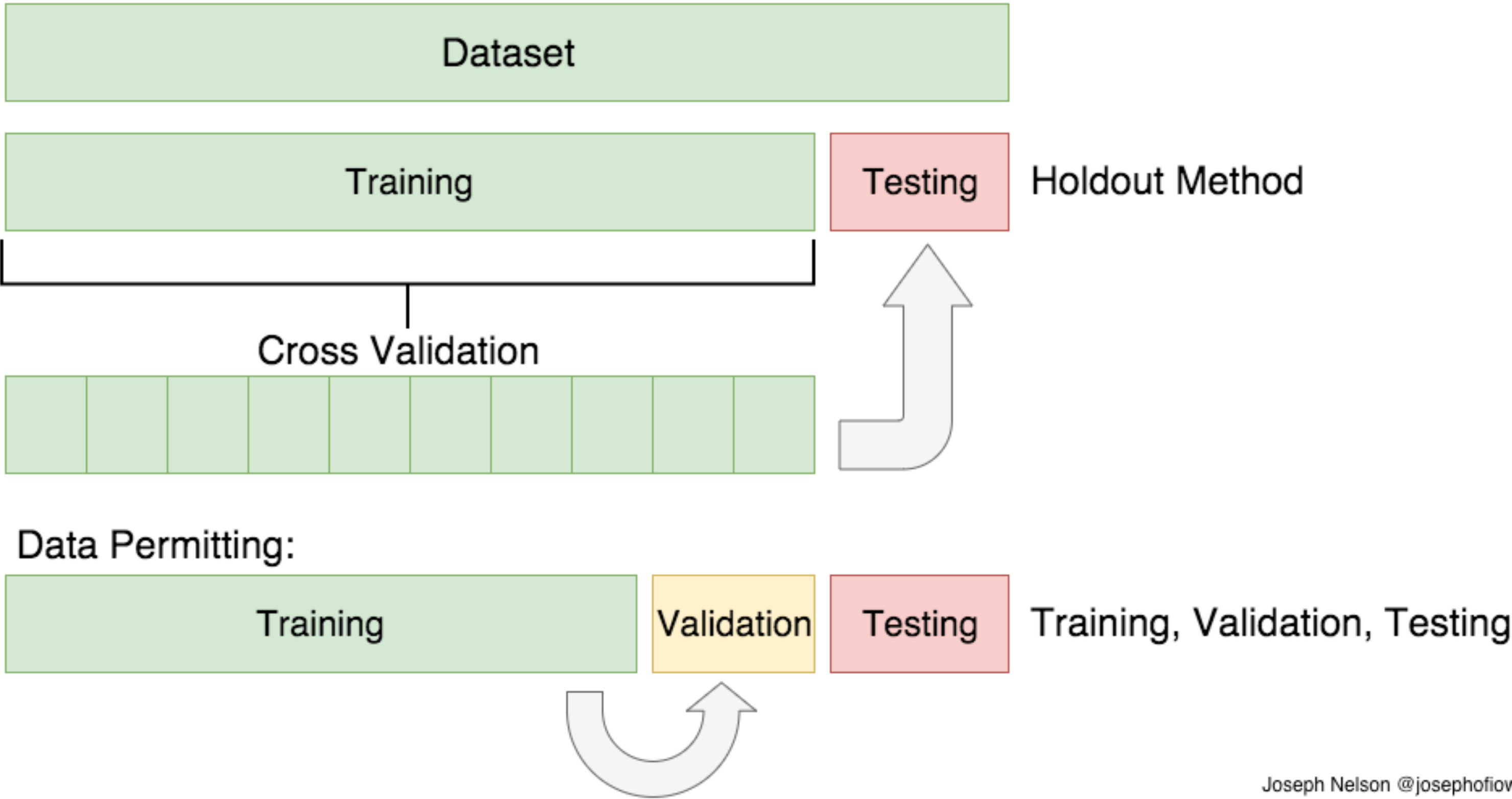
7조  
**고정욱, 편설인, 이원재**

2조  
**배광빈, 송세현, 공명구**

4조  
**노범용, 유정오, 양영규, 김영법**

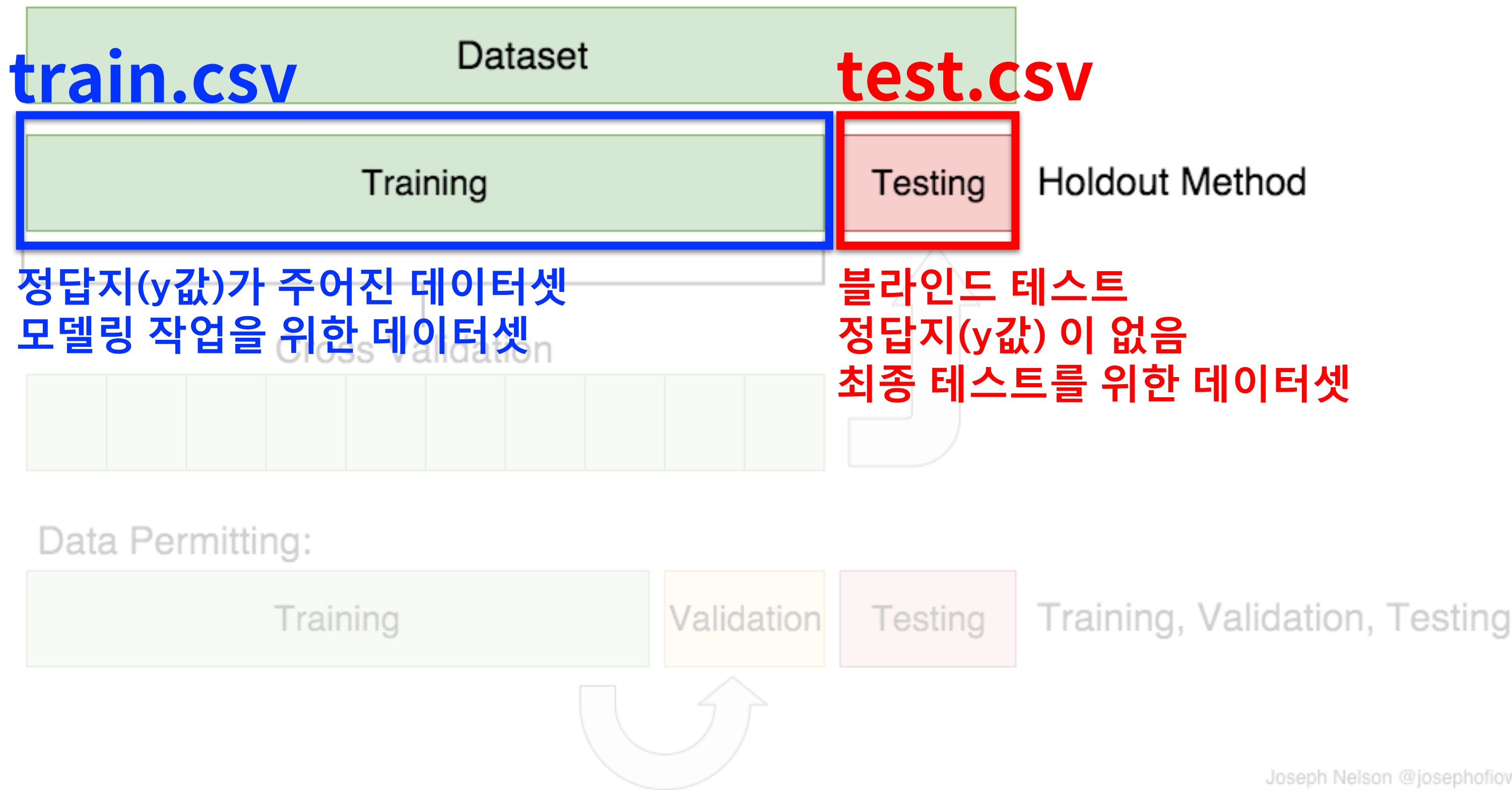
6조  
**최윤철, 이기훈, 김성희, 안동순**

데이터셋 소개

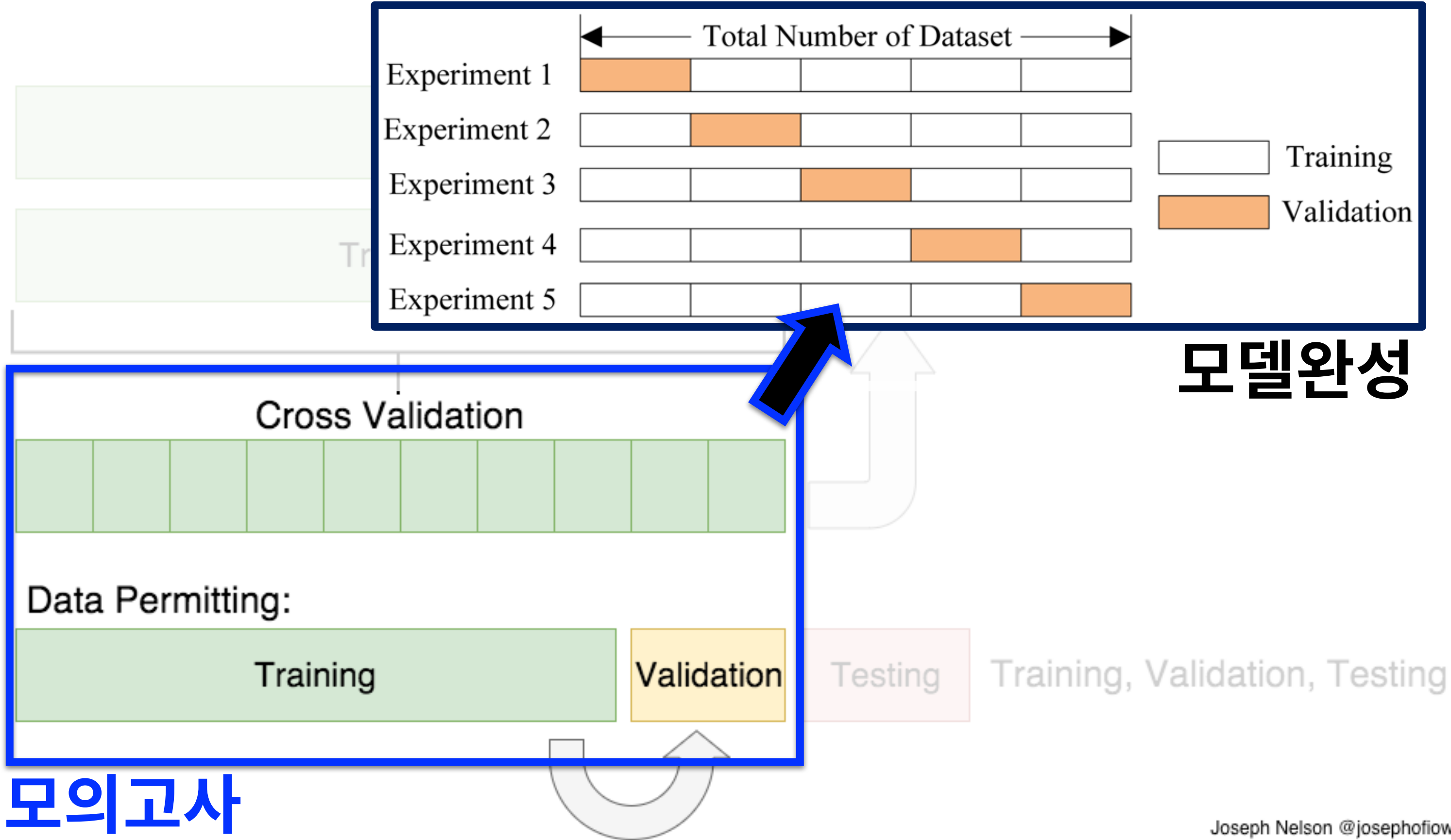


DATA SCIENCE SCHOOL

데이터셋 소개 : Train vs Test

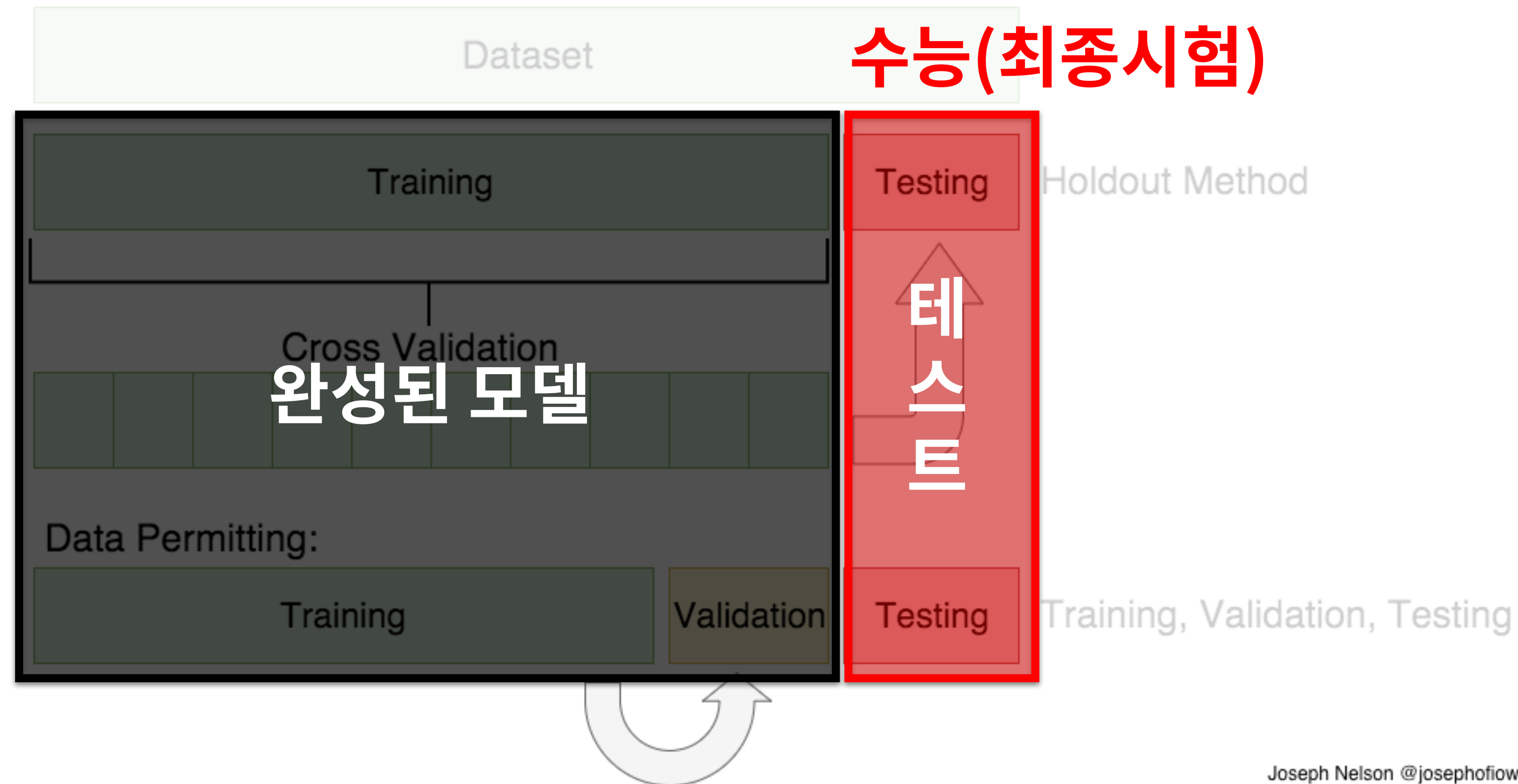


데이터셋 소개 : Cross Validation



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.)

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

## 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 SCIENCE SCHOOL

### DATA SIZE

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



## 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

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

$\epsilon$  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, 11columns (@dropoff\_datetime, trip\_duration)
- test.csv : 346797rows, 9columns



## 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

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

$\epsilon$  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, 292columns (@price\_doc)
- test.csv : 8901rows, 291columns
- macro.csv : 2484rows, 100columns - 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.)

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



# 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, 39columns (@Price)
- test.csv : 417rows, 38columns

## 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**

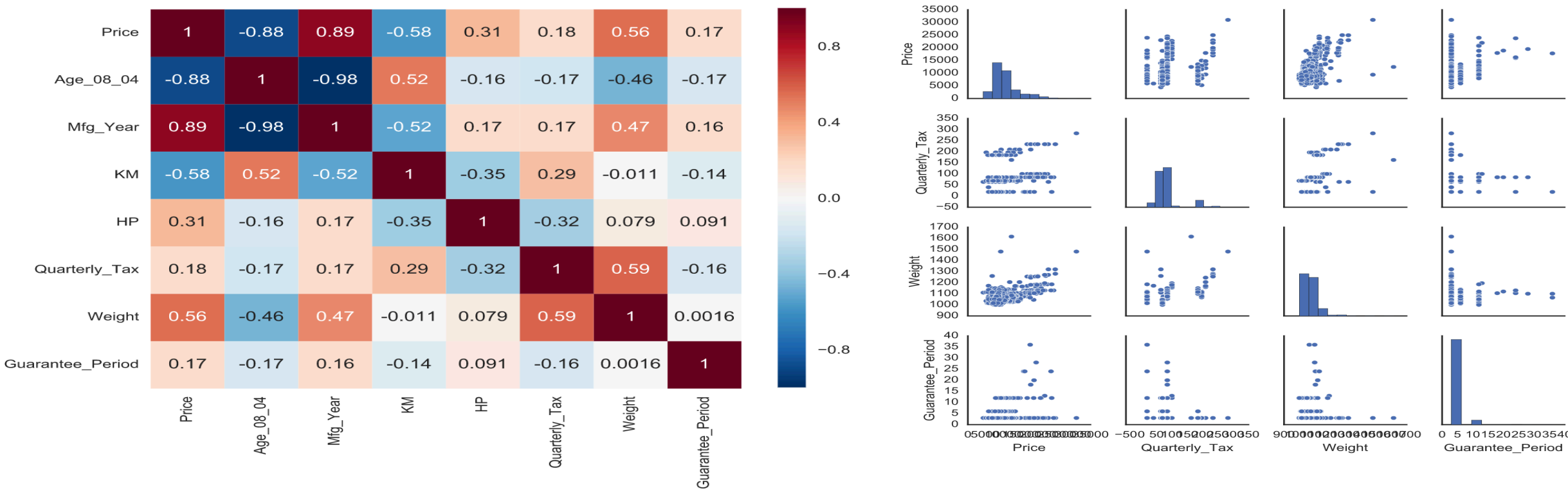
github.com/JKeun/dss-regression-datasets

DATA SCIENCE SCHOOL

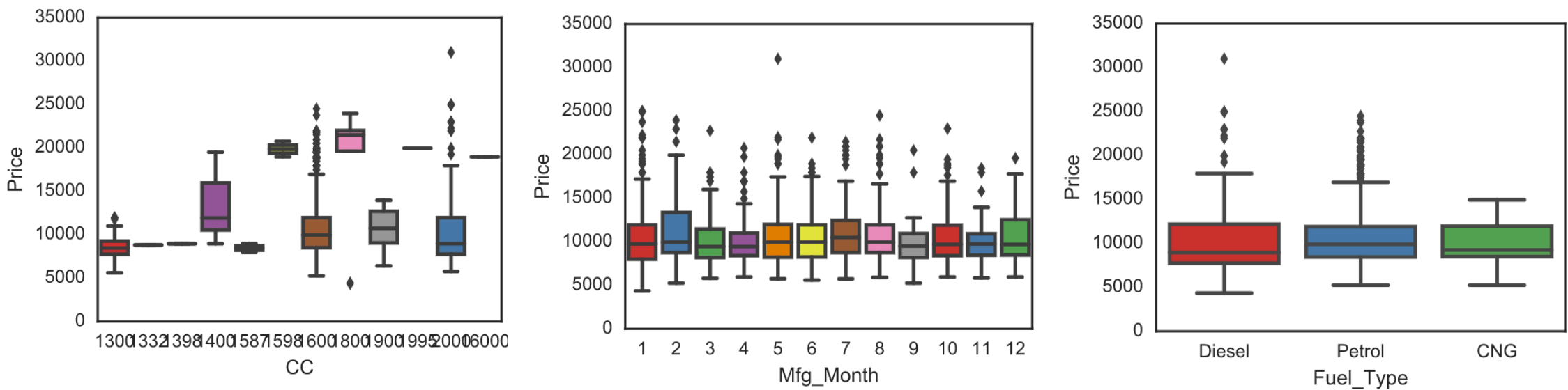
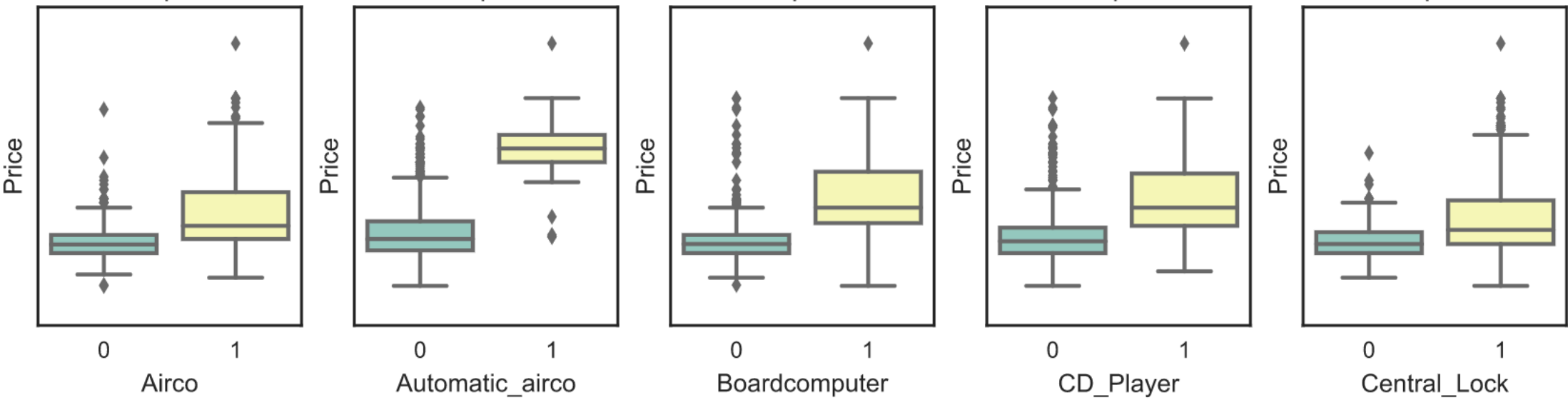
타임라인

~10/10

EDA  
Exploratory data analysis



t : -15.8161, p-val : 0.0000t : -23.6140, p-val : 0.0000t : 24.2080, p-val : 0.0000t : -18.0722, p-val : 0.0000t : 12.3531, p-val : 0.0000





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										
데이터 부족 (n = 904, 106, 9)										
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 0	
3	Model		version = CC/1000 이므로 CC에 한전종속적	Model Description		0			Weight	합에 의한 결
4	Price		종속변수 y	Offer Price in EUROS		0			1	
5	Age_08_04		채택	Age in months as in	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, C	기종 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 차종 4개, 차별성 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 i	환경보람금/	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	보장 기간	0	1	0.5	0.166	0.000
19	ABS		SUM_OPTION으로 통합	Anti-Lock Brake Sy	자동 브레이크	1	1	0.5	0.3	
20	Airbag_1		SUM_OPTION으로 통합	Driver_Airbag (Yes)	운전자 에어백	1	1	0.5	0.096	
21	Airbag_2		SUM_OPTION으로 통합	Passenger Airbag	(조수석 에어백	1	1	0.5	0.24	
22	Airco		SUM_OPTION으로 통합	Airconditioning (Ye	에어컨 유무	1	1	0.5	0.44	
23	Automatic_airco		SUM_OPTION으로 통합	Automatic Aircondit	자동 에어컨	1	1	0.5	0.6	
24	Boardcomputer		SUM_OPTION으로 통합	Boardcomputer (Ve	터스트 미터 컴퓨터 유무	1	1	0.5	0.5	

타임라인

~10/16

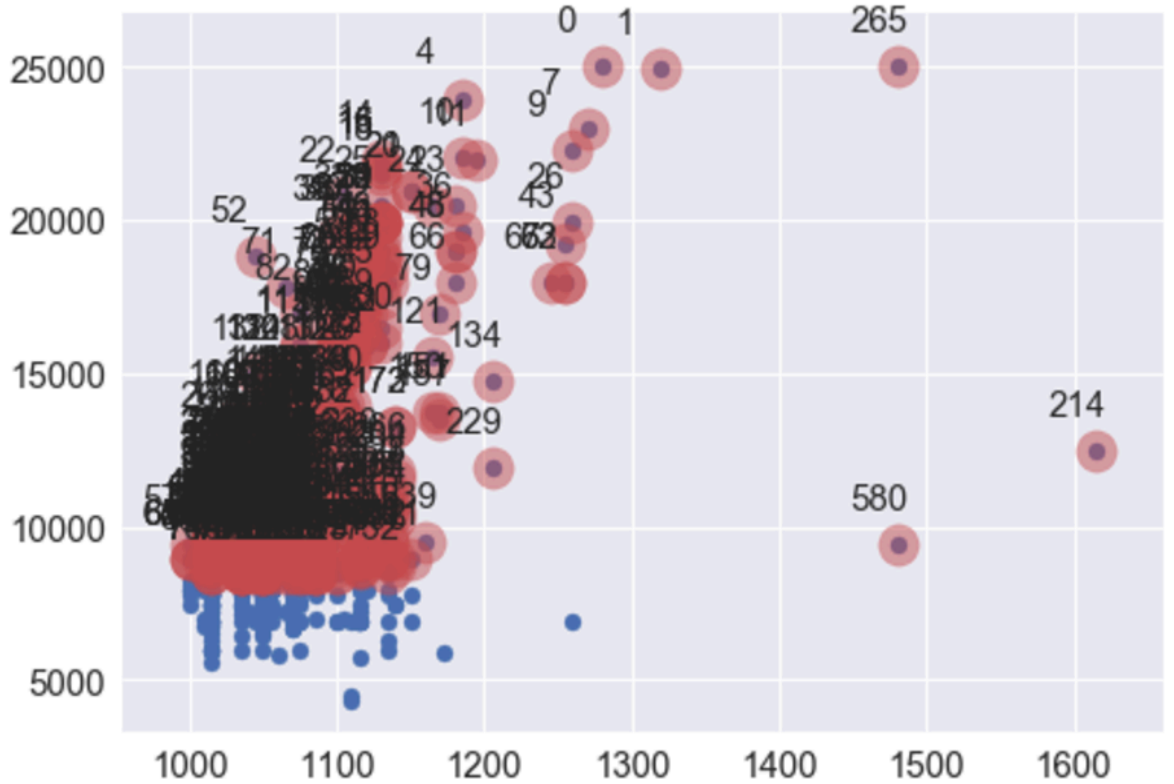
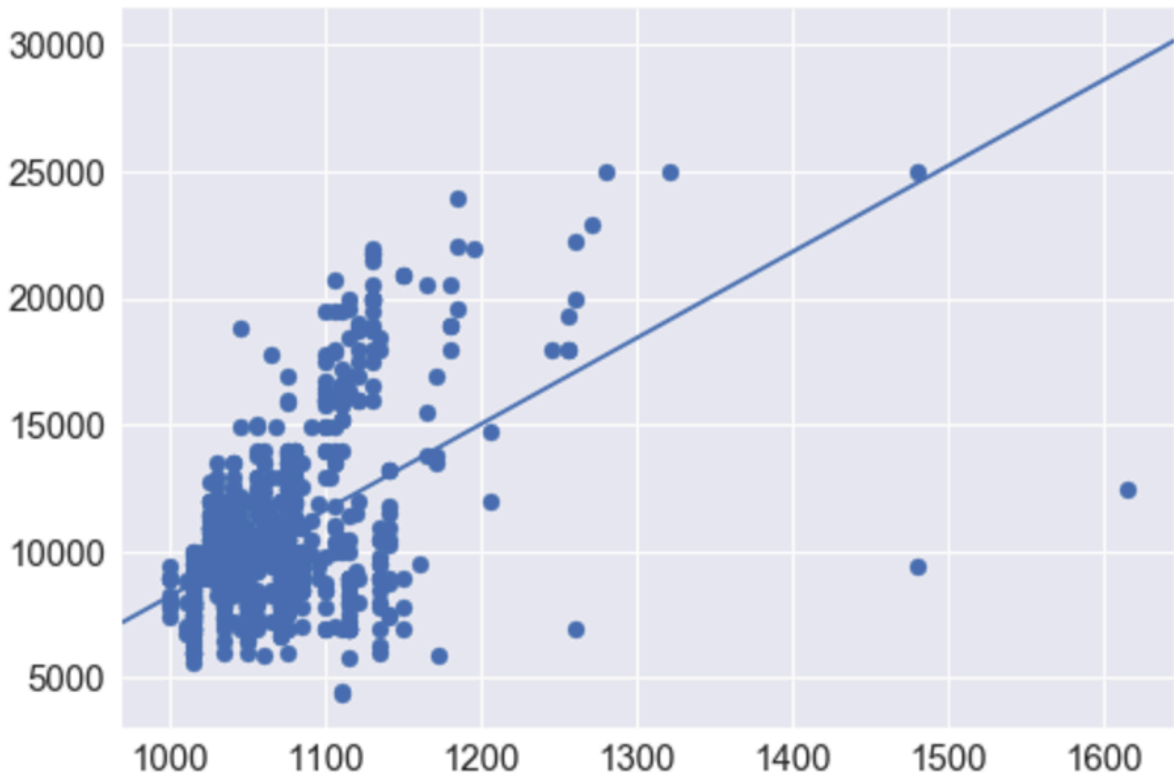
MODELING

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

$$\text{Price} = 19,230 - 138.583(\text{Age\_08\_04}) - 0.0165(\text{KM}) + 546.796(\text{InOut}) - 396.891(\text{Safe})$$

DATA SCIENCE SCHOOL

## 타임라인

~10/25

# PROJECT 발표

10/24 까지 **projectname\_answer.txt** 파일 제출

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

**10월 25일 수요일** 프로젝트 발표



END

THANK YOU ; - )

DATA SCIENCE SCHOOL