

**Kaggle**

**Data handling**

**Director of TEAMLAB  
Sungchul Choi**



# Kaggle Challenge

kaggle<sup>TM</sup>

# Data analysis Competition

---

# Google is acquiring data science community Kaggle

Posted Mar 7, 2017 by [Frederic Lardinois \(@fredericl\)](#), [Matthew Lynley \(@mattlynley\)](#), [John Mannes \(@JohnMannes\)](#)



AdChoices

## Crunchbase

Kaggle		—	
FOUNDED			
2010			
OVERVIEW			

**기업은 데이터를  
분석가는 분석기법을**

# Titanic



Getting Started Prediction Competition

## Titanic: Machine Learning from Disaster

Start here! Predict survival on the Titanic and get familiar with ML basics



Kaggle · 8,582 teams · 2 years to go



## House Prices: Advanced Regression Techniques

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

1,878 teams · 2 years to go

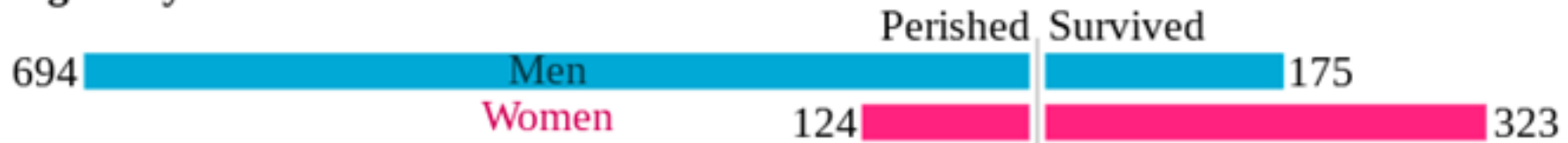


# Titanic Project

<https://www.kaggle.com/c/titanic>

- Titanic에 탑승한 승객 정보를 승객의 구출 여부를 결정

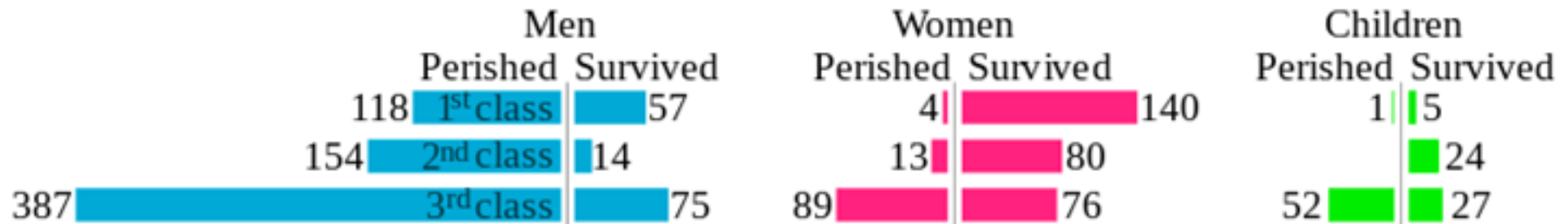
Passengers by Gender



Crew by Gender



Adult Passengers by Gender and Class, Children by Class



# Titanic Project

<https://www.kaggle.com/c/titanic>

## Variable

## Definition

survival

Survival

pclass

Ticket class

sex

Sex

Age

Age in years

sibsp

# of siblings / spouses aboard the Titanic

parch

# of parents / children aboard the Titanic

ticket

Ticket number

fare

Passenger fare

cabin

Cabin number

embarked

Port of Embarkation

## Key

0 = No, 1 = Yes

1 = 1st, 2 = 2nd, 3 = 3rd

Sibling = brother, sister, stepbrother, stepsister

Spouse = husband, wife (mistresses and fiancés were ignored)

Parent = mother, father

Child = daughter, son, stepdaughter, stepson

Some children travelled only with a nanny, therefore parch=0 for them.


C = Cherbourg, Q = Queenstown, S = Southampton

<https://www.kaggle.com/c/titanic/data>

# Titanic Project

<https://www.kaggle.com/c/titanic>

- Data는 testset과 training set을 제공
- Testset으로 모델을 만든 후 trainset에 적용
- 결과제출은 [ID , 생존 예측] 형태로 제출
- 제출된 결과를 바탕으로 accuracy 점수로 등수를 산정함
- 분석가들은 기존 자신들이 시도했던 다양한 분석 방법을 사이트를 통해서 공유하고 있음

 gender\_submission.cs...

 test.csv









 train.csv

**Public Leaderboard****Private Leaderboard**

















This leaderboard is calculated with approximately 50% of the test data.

The final results will be based on the other 50%, so the final standings may be different.

[Raw Data](#)[Refresh](#)

#	Δ1w	Team Name	Kernel	Team Members	Score ?	Entries	Last
1	—	<b>Nikhil Mathew</b>			1.00000	13	2mo
2	—	<b>Shahnab</b>			1.00000	1	2mo
3	—	<b>MuhammadMusab</b>			1.00000	5	2mo
4	—	<b>Raevent</b>			1.00000	1	1mo
5	—	<b>Onipe theoderic</b>			1.00000	5	1mo
6	—	<b>RANDOF Consulting</b>			1.00000	2	22d
7	—	<b>ryemitan</b>			1.00000	1	21d
8	—	<b>Jason Zutty</b>			1.00000	2	20d



		Sort by Hotness ▾			
All Mine		All Languages ▾		All Output Types ▾	
▲ 58		 <b>EDA To Prediction(DieTanic)</b> run 8 hours ago by <a href="#">ashwin</a>  data visualization, classification, ensembling, eda, model comparison	Py	25 	
▲ 1		<b>Titanic Prediction thru. ML</b> run 4 hours ago by <a href="#">RaviKaushik</a>  data analysis, beginner, random forest, feature engineering	R		
▲ 1		<b>My Titanic Analysis</b> run 35 minutes ago by <a href="#">Tim Leung</a> (+164 / -716)	Py		
▲ 6		<b>Finding Important Factors To Survive Titanic</b> run 10 hours ago by <a href="#">Jatturat Janejarasskul</a>  classification, data visualization, data analysis, feature engineering	Py	4 	
▲ 0		<b>Predicting Survival -CART &amp; Ensemble Methods</b> run an hour ago by <a href="#">Pamela Augustine</a>	Py		
▲ 31		 <b>Random forest+Logistic.R+Decison.T+SVM for Titanic</b> run 21 hours ago by <a href="#">swamysm</a> (+525 / -385)  logistic regression, svm, decision tree, random forest	R	18 	
▲ 0		<b>titanic_try_20171017 [yeah! 0.80382]</b> run 4 hours ago by <a href="#">sion.</a>	Py		

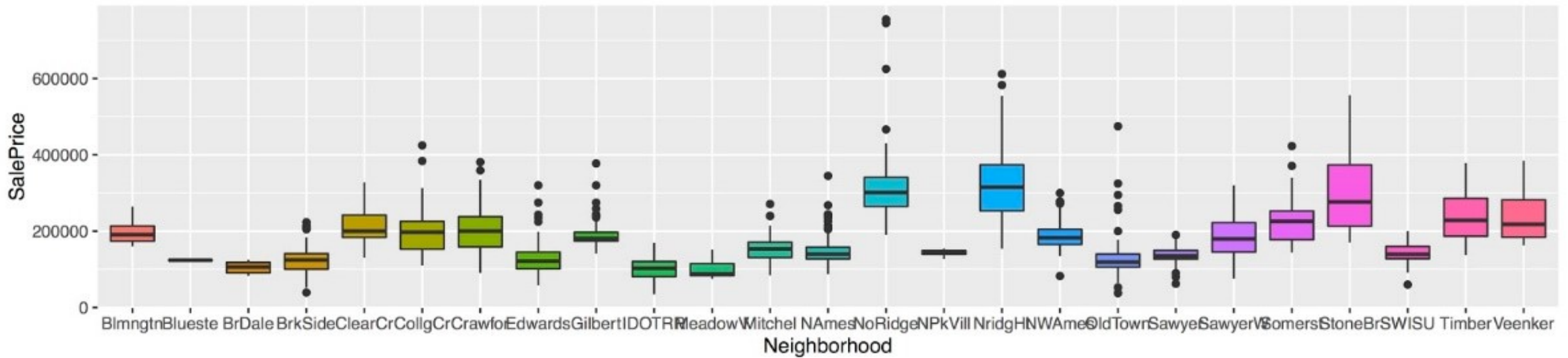
# Challenge

<https://www.kaggle.com/ash316/eda-to-prediction-dietanic>

House price

# Advance House Price

- 기존 Boston House Price의 Advance 문제
- 80여개의 Feature를 Handling 하는 연습 문제



<https://www.kaggle.com/c/house-prices-advanced-regression-techniques>



---

# Advance House Price

- **SalePrice** - the property's sale price in dollars. This is the target variable that you're trying to predict.
- **MSSubClass**: The building class
- **MSZoning**: The general zoning classification
- **LotFrontage**: Linear feet of street connected to property
- **LotArea**: Lot size in square feet
- **Street**: Type of road access
- **Alley**: Type of alley access
- **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 main road or railroad
- **Condition2**: Proximity to main road or railroad (if a second is present)
- **BldgType**: Type of dwelling
- **HouseStyle**: Style of dwelling
- **OverallQual**: Overall material and finish quality
- **OverallCond**: Overall condition rating
- **YearBuilt**: Original construction date
- **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
- **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**: Number of bedrooms above basement level
- **Kitchen**: Number of kitchens

# Advance House Price

- 어떤 데이터는 한 개가 여러개로 분리되어있음 (방개수?)
- 데이터가 너무 많아 Drop 해야하는 데이터가 많음
- 데이터에 대한 이해가 필요
- 일단 그냥 전처리해서 모델 만들기도 쉽지 않음
- Numeric Value와 Category Value를 나눠서 접근!!

# Challenge