# DiSCo

## BGU's Data Science Community

MEET • TEAM UP • KAGGLE

HAVE FUN!

DiSCo
Data Science Community

Bengis Center for Entrepreneurship & Innovation
Guilford Glazer Faculty of Business and Management
Ben-Gurion University of the Negev

# Where we are?

- How to use Kaggle ?
  - Data description
  - Making kernel
  - Submitting predictions
- Started with “House prices competition” focused on Advanced Regression
- Overview on preprocessing also known as data cleaning
- Feature engineering examples
- Making predictions with advanced models including
  - Linear Regression
  - Tree boosting
  - Deep learning

# What now?

- Importance of each feature
  - Look deep
  - Are you going to bring? Or just making noise?
- Relation between features?
  - Don’t treat me as just data
- Can you create better feature out of existing one?
  - How that magic work?
- What model to use?
  - Relevant to problem
  - Understand architecture

kaggle Search kaggle Competitions Datasets Kernels Discussion Jobs


**House Prices: Advanced Regression Techniques**

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

4,375 teams · 2 years to go

Overview Data Kernels Discussion Leaderboard Rules Team My Submissions Submit Predictions

- Description
- Evaluation
- Frequently Asked Questions
- Tutorials

**Start here if...**

You have some experience with R or Python and machine learning basics. This is a perfect competition for data science students who have completed an online course in machine learning and are looking to expand their skill set before trying a featured competition.

**Competition Description**

Ask a home buyer to describe their dream house, and they probably won't begin with the height of the basement ceiling or the proximity to an east-west railroad. But this playground competition's dataset proves that much more influences price negotiations than the number of bedrooms or a white-picket fence.

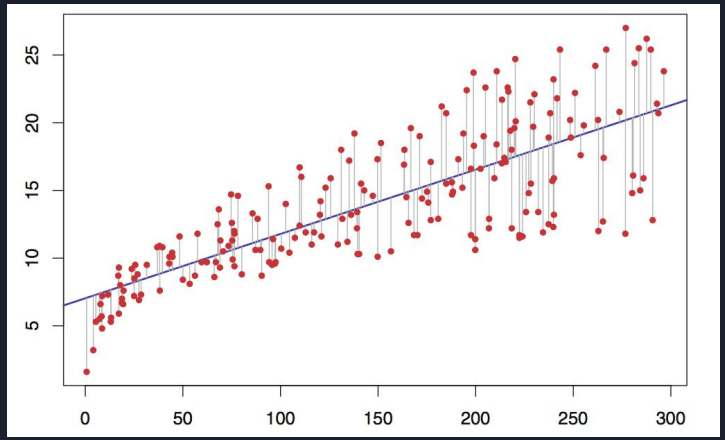
With 79 explanatory variables describing (almost) every aspect of residential homes in Ames, Iowa, this competition challenges you to predict the final price of each home.

# House Prices Competition

# Linear regression : Mean Squared Error

Euclidean distance between the target and predicted.

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \tilde{y}_i)^2$$




$$E(W) = \frac{1}{2}\sum_{i=0}^{n}(W.X_i^T - Y_i)^2$$

MSE(Mean Squared Error) and Error Function

# Linear regression : gradient descent

$$E(W) = \frac{1}{2}\sum_{i=0}^{n}(W.X_i^T - Y_i)^2$$

*The least squared loss of a linear model is a convex function (“bowl-shaped”)*

One simple way to find its minimum is by **following the slope of the error**.

$$W \leftarrow W + \alpha\frac{\delta E}{\delta W}$$

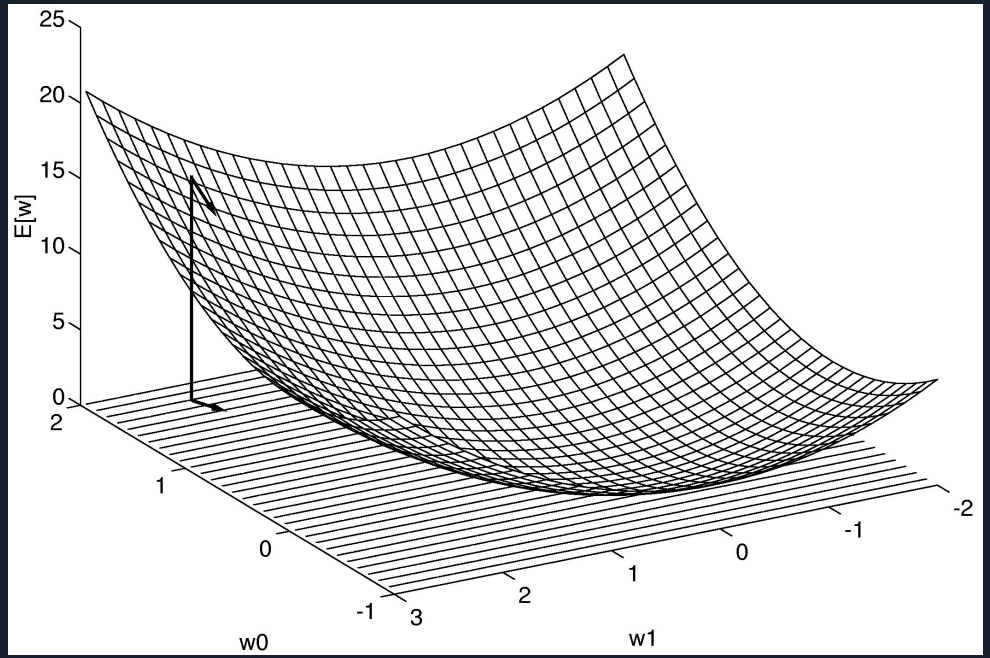
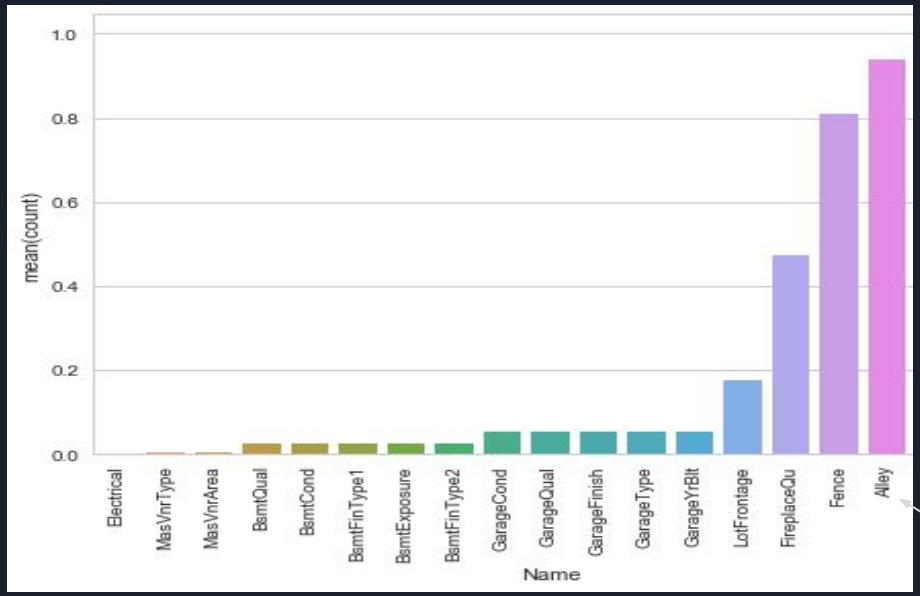
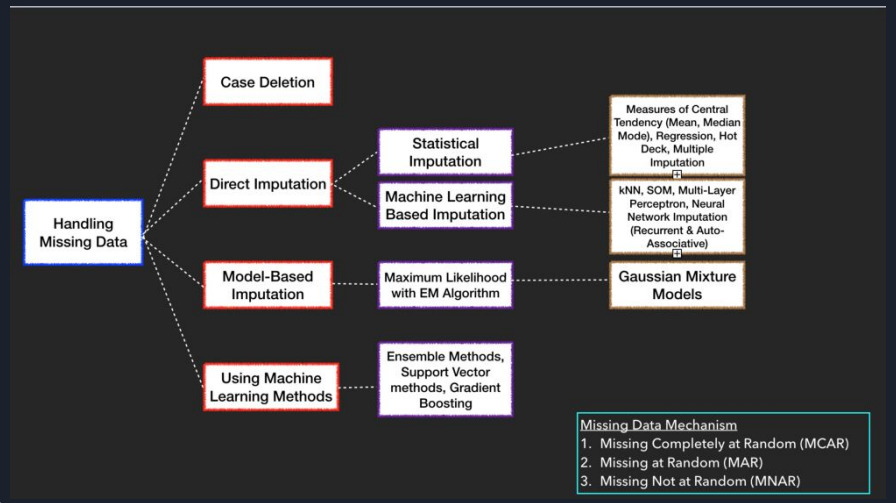



*Illustration: Tom Mitchell, McGraw-Hill*

# Univariate Analysis - Plot the missing values count for features

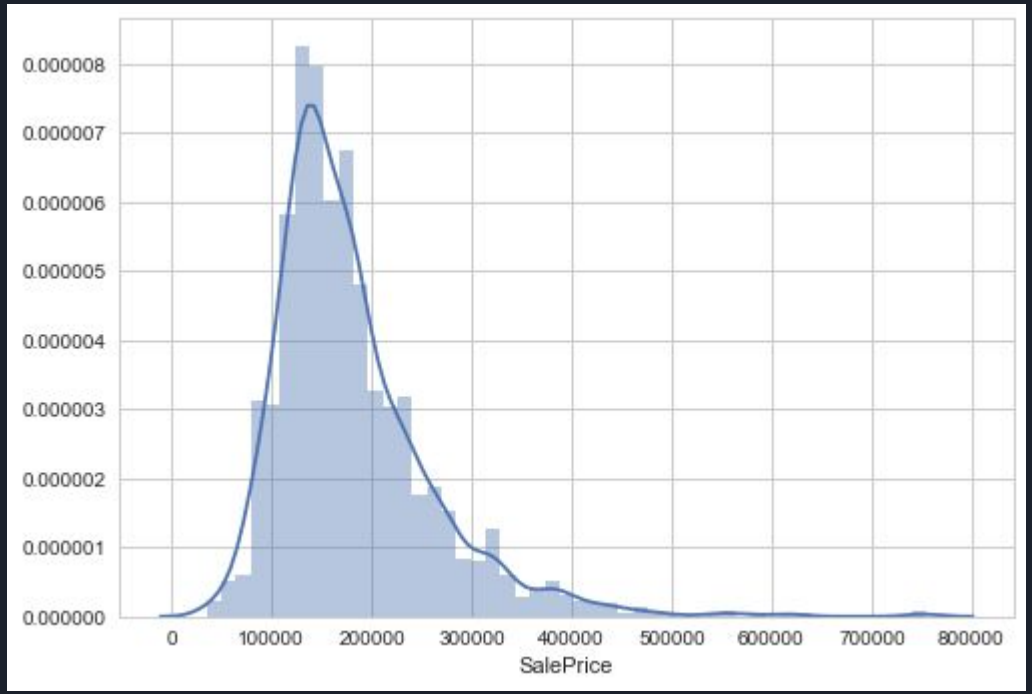


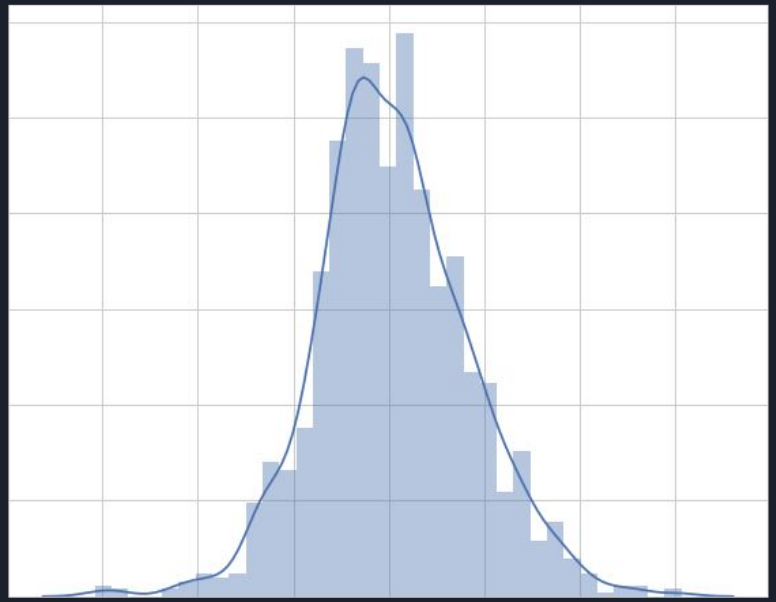

| Id | MSSubClass | MSZoning | LotFrontage | LotArea | Street | Alley | LotShape | LandContour | Utilities |
|---|---|---|---|---|---|---|---|---|---|
| 1 | 60 | RL | 65.0 | 8450 | Pave | NaN | Reg | Lvl | AllPub |
| 2 | 20 | RL | 80.0 | 9600 | Pave | NaN | Reg | Lvl | AllPub |
| 3 | 60 | RL | 68.0 | 11250 | Pave | NaN | IR1 | Lvl | AllPub |
| 4 | 70 | RL | 60.0 | 9550 | Pave | NaN | IR1 | Lvl | AllPub |
| 5 | 60 | RL | 84.0 | 14260 | Pave | NaN | IR1 | Lvl | AllPub |

# Univariate Analysis - Transformation of 'Target Variable' - Sales Prices

- Skewness : A measure of assymetry in the distribution
- SalesPrices distribution is concentrated to the left

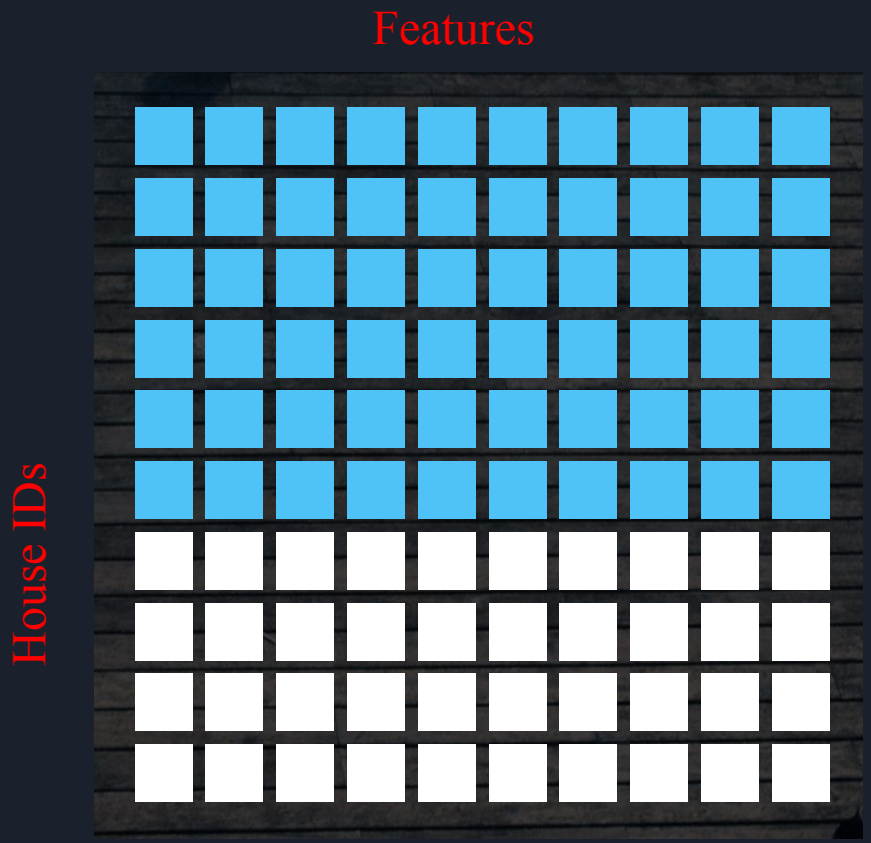



**Positively Skewed distribution**



**Normal Distribution**

# ML workflow




Training Set

Test Set

Model Building using training set

Cross Validation

Prediction using test set

# Acknowledgement


## Course Creation and people responsible

Administration, Fundraising , Public connection - Moran Sharon & Ruth Hashkes

Github page, kernels for week 3 - Minesh Jethva

Making the presentation and management support - Rahul Veettil


Thank you !

# Ok! Let's get our hands dirty

**From Week 1**

**https://discobgu.github.io/**

**https://tinyurl.com/disco-kernel1**
**https://tinyurl.com/disco-facebook**

**For Week 2**

**https://tinyurl.com/kernal-week2**
**https://tinyurl.com/week2-resources**

**Week 3**
**https://tinyurl.com/week3-resourses**

1. Minimal Kernel LB: 0.60109
   - NaN => Median
   - LinearRegression
2. Minimal + Normalized X Kernel LB: 0.30013
   - LinearRegression(Normalized X)
3. Minimal + Normalized X,y Kernel LB: 0.14305
   - y = log2(y)
4. Minimal + Normalized X skew,y Kernel LB: 0.14104
   - X = log2(X) if abs(skew) > 1.7 & no Inf issues
5. Minimal + Normalized X skew,y + filter low Var Kernel LB: 0.13764
   - filter X if Variance < 0.2 and not correlated with target y

# The DiSCo Team

**Rahul Veettil**

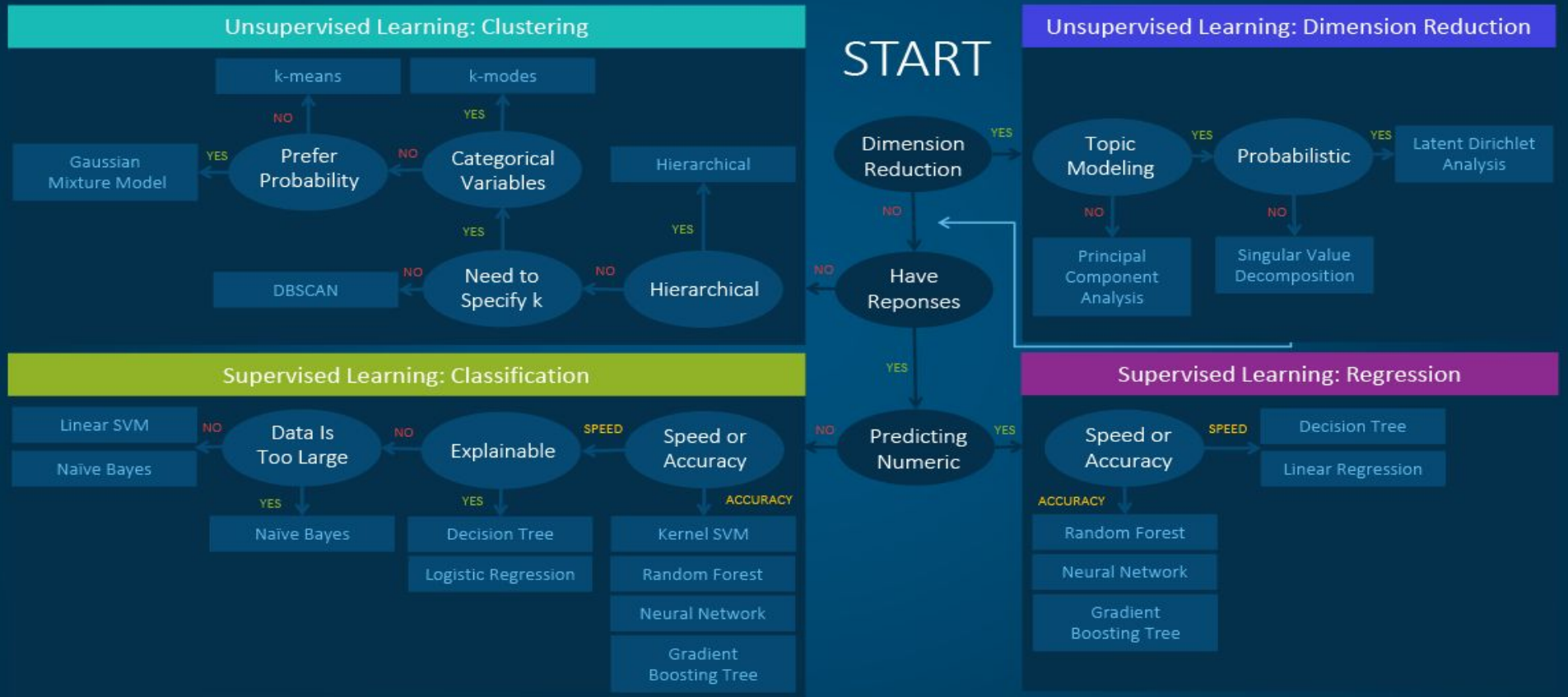
**Minesh Jethva**

**Ruth Hashkes**

**Moran Sharon**

https://www.bengis.org/disco
https://discobgu.github.io
Contact : disco.bgu@gmail.com

# Machine Learning Algorithms Cheat Sheet




https://blogs.sas.com/content/subconsciousmusings/2017/04/12/machine-learning-algorithm-use/?utm_source=facebook&utm_medium=cpc&utm_campaign=analytics-global&utm_content=US_interests-conversions