

FIRST MEETUP

Welcome to

# DiSCo

BGU's Data Science  
Community

MEET • TEAM UP • KAGGLE  
HAVE FUN!

Sign Up Here: [tinyurl.com/discobgu](http://tinyurl.com/discobgu)



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

BEN-GURION UNIVERSITY OF THE  
NEGEV

# 25 MARCH 2018

DOORS OPEN AT 18:30

- 18:30-19:00 - WELCOME AND INTRODUCTION. WHO ARE WE AND WHAT ARE THE COMMUNITY'S GOALS.
- 19:00-20:00 - LET THE FUN BEGIN! CREATING GROUPS AND INITIATING THE FIRST KAGGLE COMPETITION.



Learning data science through kaggle

# Team DiSCo



Rahul Veettill



Minesh Jethva



Ruth Hashkes



Moran Sharon

<https://www.bengis.org/disco>

# Are Data Scientists Prophets of the Modern Era?

25th-March-2018 - Week 1

DiSCo - Data Science Community  
Learning data science through kaggle

# Predictions by Nostradamus



Death of the  
Kennedy Brothers

---

Reign of Hitler

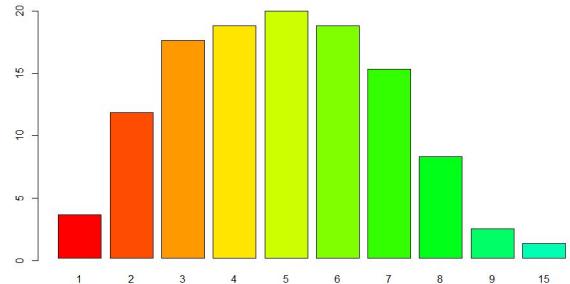
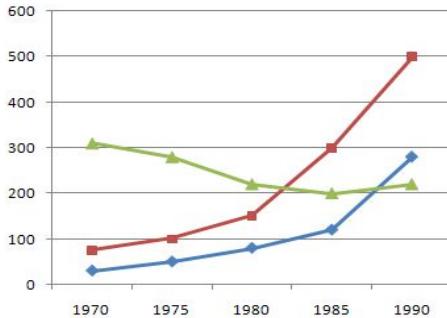
---

The  
French Revolution

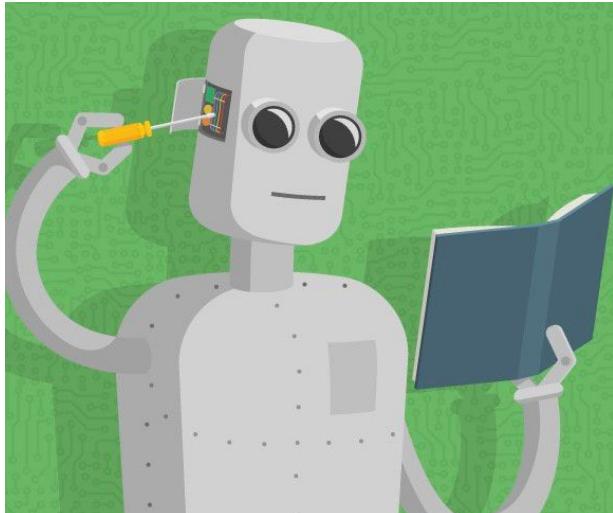
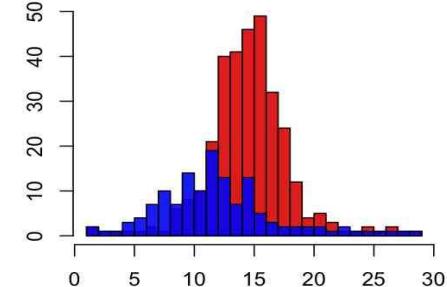
---

9/11

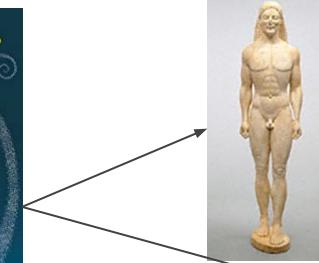
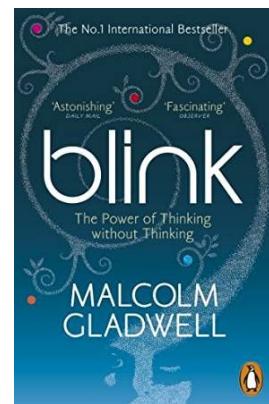
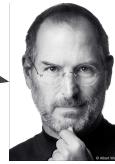
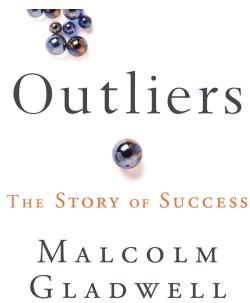
# Agenda



python



# The ‘science’ in Data Science



# The intriguing story of Heart Attack Diagnosis

**Algorithm was 70 percent better than the old method**

Cook County Hospital

2018



1970s



## Problem Statement

How to decide whether a person in the emergency room is having a heart attack and how serious it is

## Solution

- Electrocardiogram reading
- Fluid in patient's lung
- Unstable angina

# Developing a data scientist's mind set

## MODERN DATA SCIENTIST

Data Scientist, the sexiest job of 21th century requires a mixture of multidisciplinary skills ranging from an intersection of mathematics, statistics, computer science, communication and business. Finding a data scientist is hard. Finding people who understand who a data scientist is, is equally hard. So here is a little cheat sheet on who the modern data scientist really is.

### MATH & STATISTICS

- ★ Machine learning
- ★ Statistical modeling
- ★ Experiment design
- ★ Bayesian inference
- ★ Supervised learning: decision trees, random forests, logistic regression
- ★ Unsupervised learning: clustering, dimensionality reduction
- ★ Optimization: gradient descent and variants

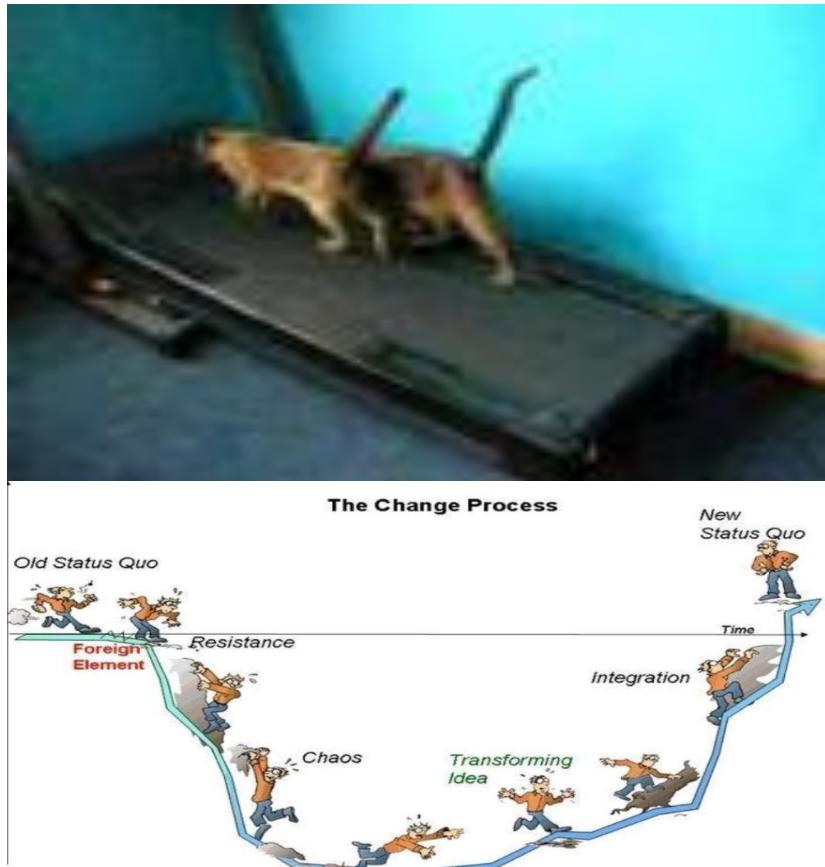


### PROGRAMMING & DATABASE

- ★ Computer science fundamentals
- ★ Scripting language e.g. Python
- ★ Statistical computing package e.g. R
- ★ Databases SQL and NoSQL
- ★ Relational algebra
- ★ Parallel databases and parallel query processing
- ★ MapReduce concepts
- ★ Hadoop and Hive/Pig
- ★ Custom reducers
- ★ Experience with xaaS like AWS

### COMMUNICATION & VISUALIZATION

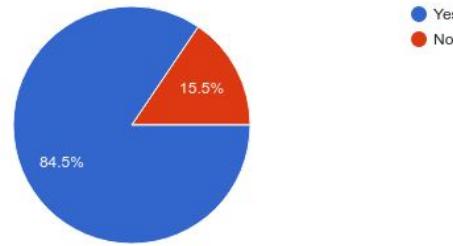
- ★ Able to engage with senior management
- ★ Story telling skills
- ★ Translate data-driven insights into decisions and actions
- ★ Visual art design
- ★ R packages like ggplot or lattice
- ★ Knowledge of any of visualization tools e.g. Flare, D3.js, Tableau



# DiSCo - Google form submission data

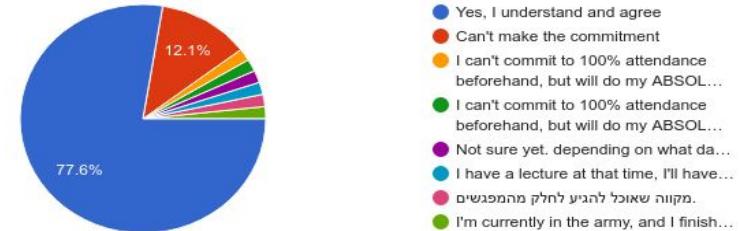
Are you a student?

58 responses



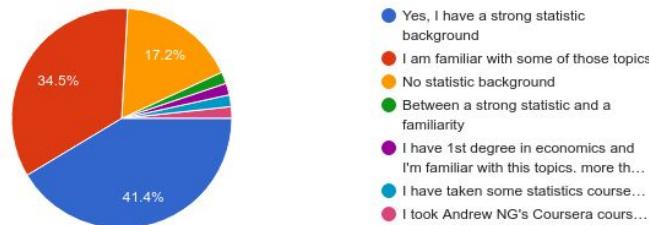
In order to create the community, we first limit the number of participants. Therefore, by signing in, you express ... community during the whole semester.

58 responses



Have you taken any courses in statistics in the past (t-test, chi-square, ANOVA, Linear regression etc)?

58 responses



Do you know programming in any language (R, Python or any other)?

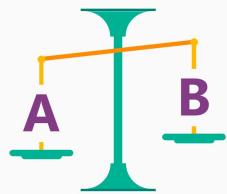
58 responses



# Introduction to Machine Learning Algorithms

Is this A or B?

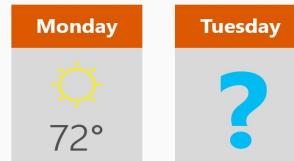
Classification algorithms



Classification

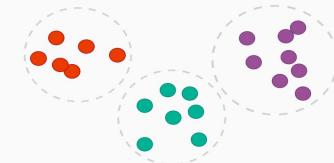
How much? How many?

Regression algorithms



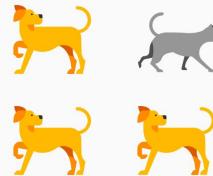
How is this organized?

Clustering Algorithms



Is this weird?

Anomaly detection algorithms



Anomaly detection

What should I do now?

Reinforcement Learning Algorithms



Reinforcement learning

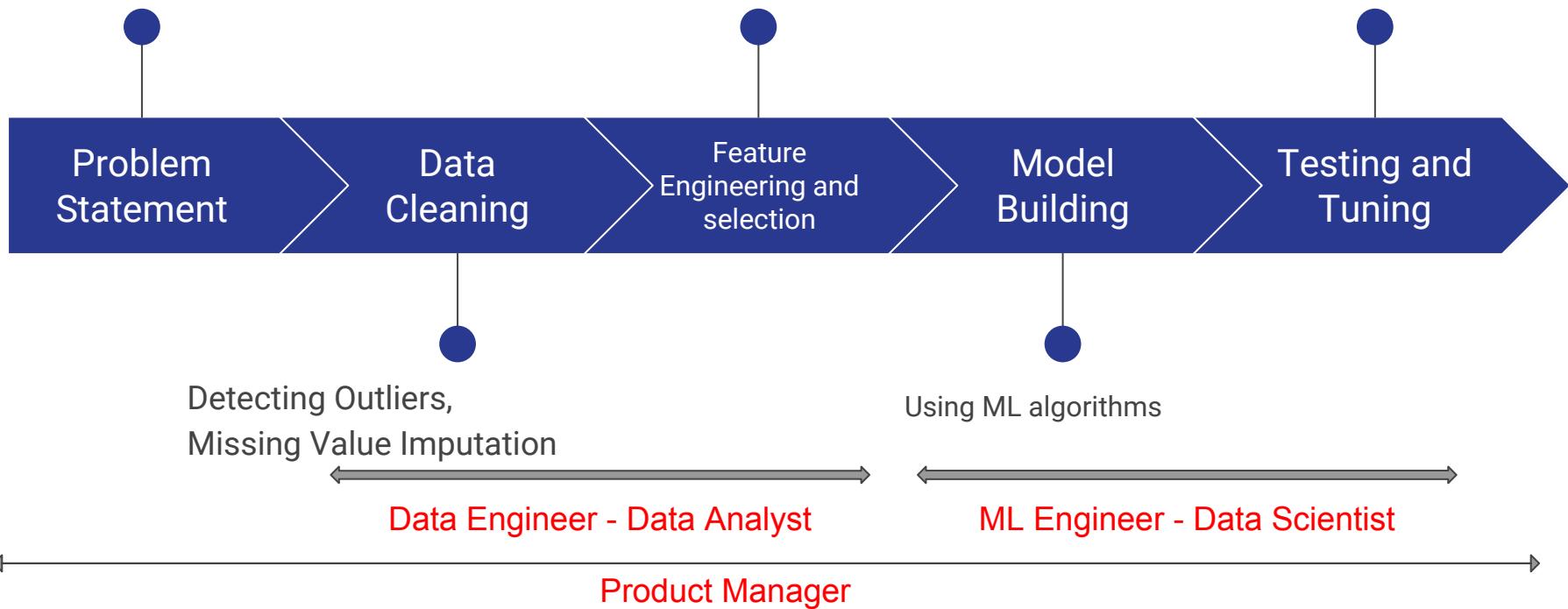


# Workflow

Determine House Sale Price

Creating, Transforming and Engineering features

Testing Model Accuracy & Parameter Optimization



# Ok! Let's get our hands dirty

<https://tinyurl.com/disco-git>

<https://tinyurl.com/disco-kernell1>

<https://tinyurl.com/disco-facebook>





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... Competition Description

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.

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

# House Price - Let's start with an exercise

<b>Square Feet</b>	<b>Sales Price</b>
100	\$ 100K
200	\$ 200K
300	\$ 300K
...	....
150	\$ ?

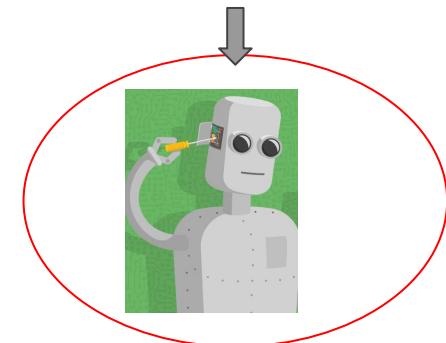
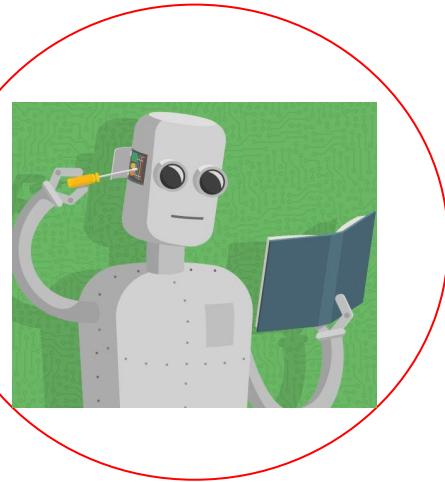
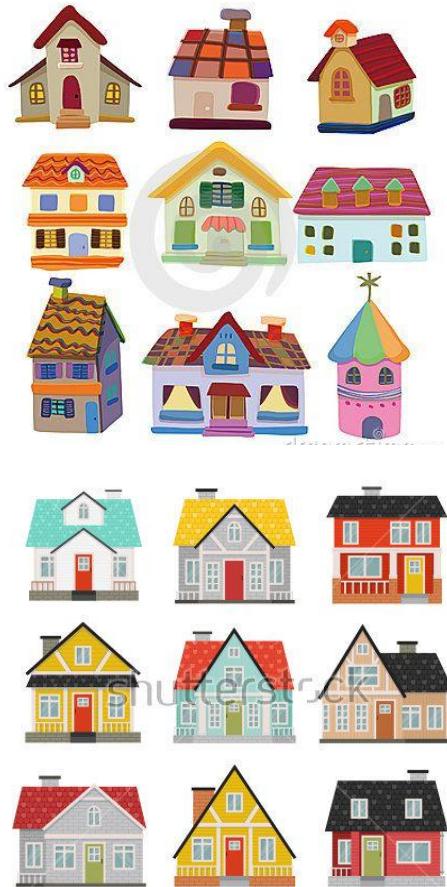
## House Price - One more exercise :)

<b>Square Feet</b>	<b>Crime Rate (no per 10k people)</b>	<b>Sales Price</b>
100	10	\$ 100K
200	40	\$ 50K
300	20	\$ 150K
...	...	....
400	30	\$ ?

## Real life scenario

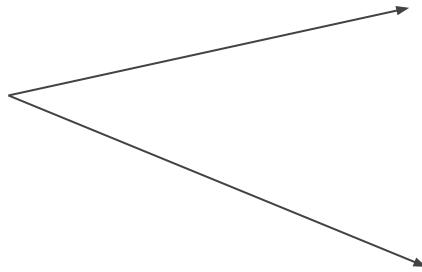
<b>Id</b>	<b>MSSubClass</b>	<b>MSZoning</b>	<b>LotFrontage</b>	<b>LotArea</b>	<b>Street</b>	<b>Alley</b>	<b>LotShape</b>	<b>LandContour</b>	<b>Utilities</b>
1	60	RL	65.0	8450	Pave	NaN	Reg	Lvl	AllPub
2	20	RL	80.0	9600	Pave	NaN	Reg	Lvl	AllPub
3	60	RL	68.0	11250	Pave	NaN	IR1	Lvl	AllPub
4	70	RL	60.0	9550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

# Problem Statement



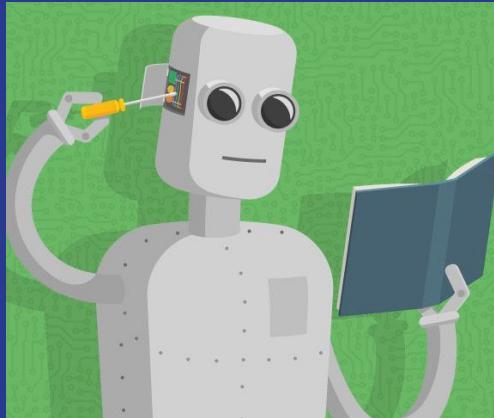
# Features of individual houses

What factors can you think of right now which can influence house prices ?



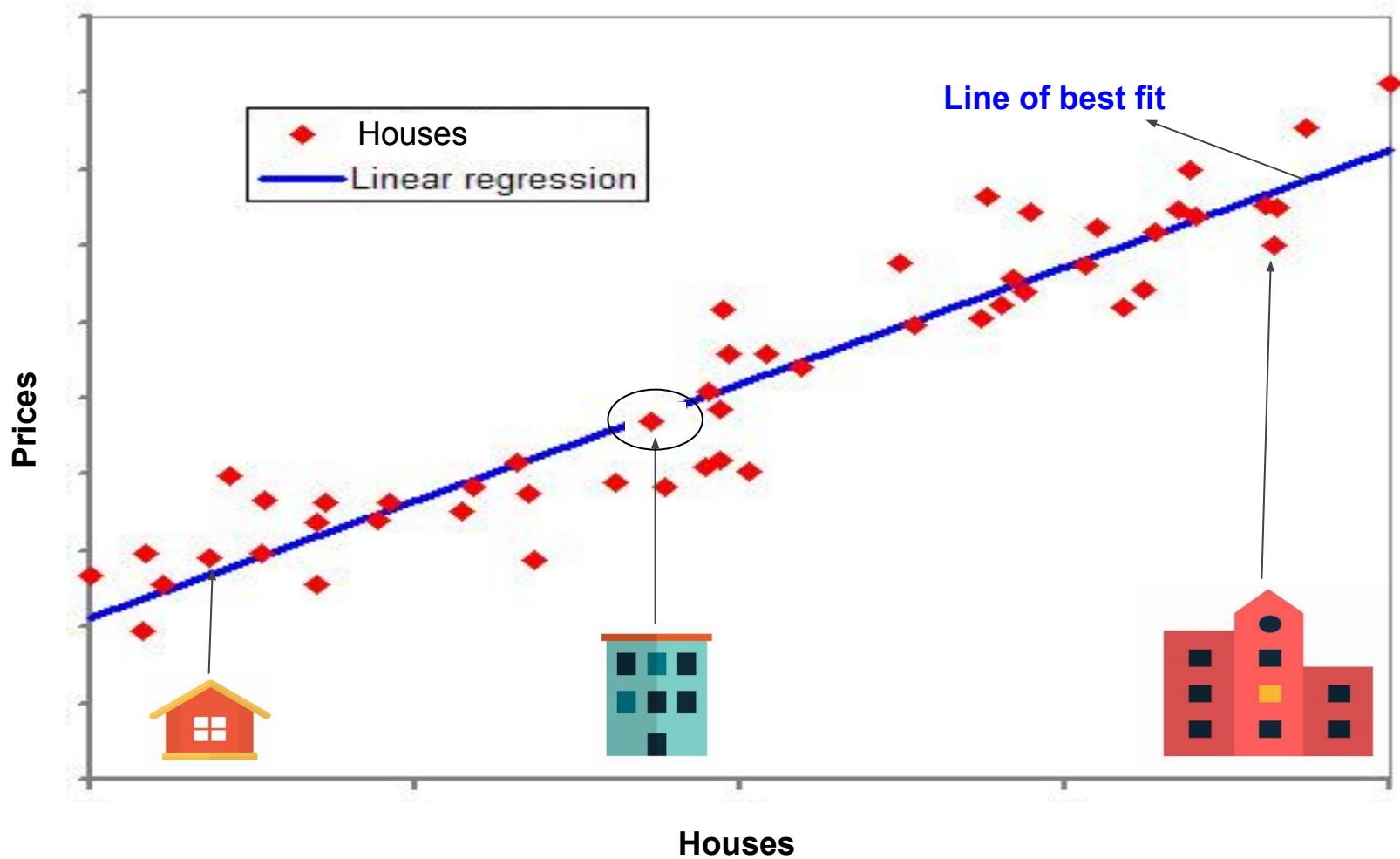
- Area of House
  - How old is the house
  - Location of the house
  - If terrace is available
  - If car parking is available
  - If security is available
- Price

# Simple Linear regression





Express-highway of  
best fit



# Linear regression : Introduction

In supervised learning, we have examples of lots of input and the desired output value.

This is called the **dataset**:

- A matrix of **feature vectors X**.
- A vector of **target values Y**.

- Area of House
- How old is the house
- Location of the house
- If terrace is available
- If car parking is available
- If security is available

- Price



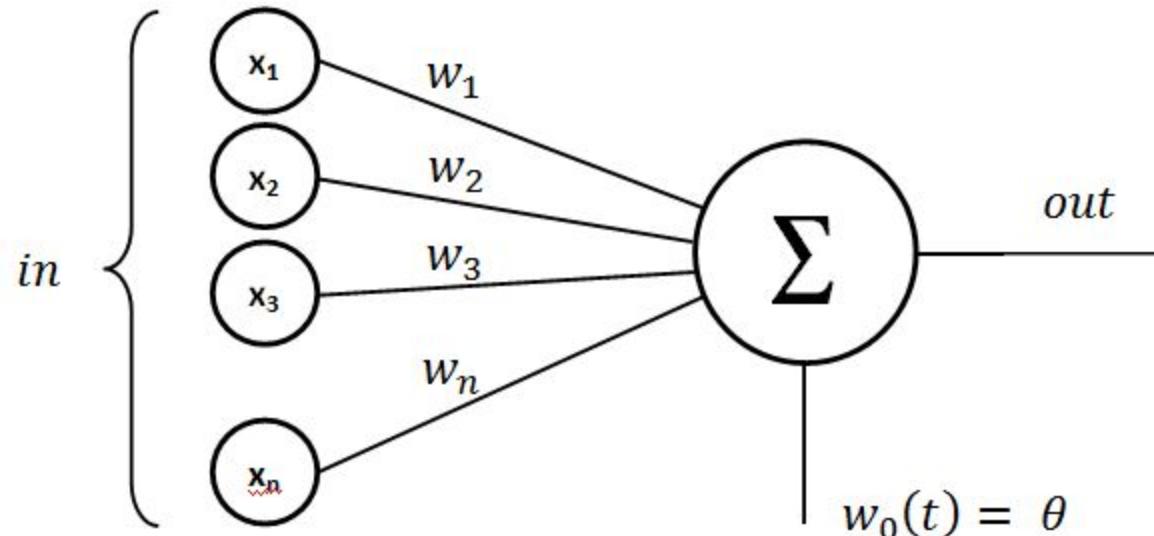
# Data Exploration

<b>Id</b>	<b>MSSubClass</b>	<b>MSZoning</b>	<b>LotFrontage</b>	<b>LotArea</b>	<b>Street</b>	<b>Alley</b>	<b>LotShape</b>	<b>LandContour</b>	<b>Utilities</b>
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

## Statistical Analysis

- 1. Univariate Analysis** - Examples: histogram, density plot, etc.
- 2. Bivariate Analysis** - Examples: bar chart, line chart, area chart, etc.
- 3. Multivariate Analysis** - Examples: stacked bar chart, dodged bar chart, etc.

# Linear model

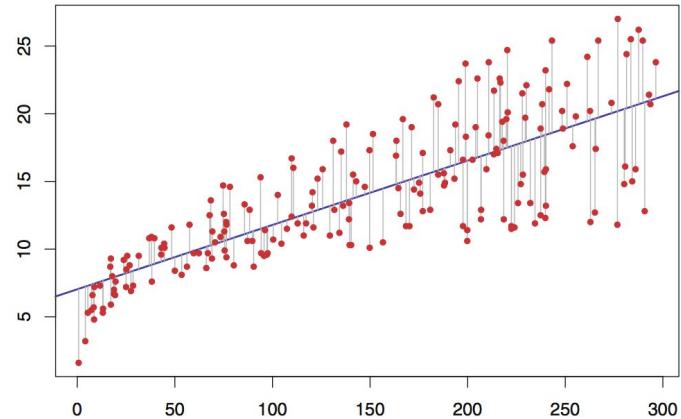


# Linear regression : error function

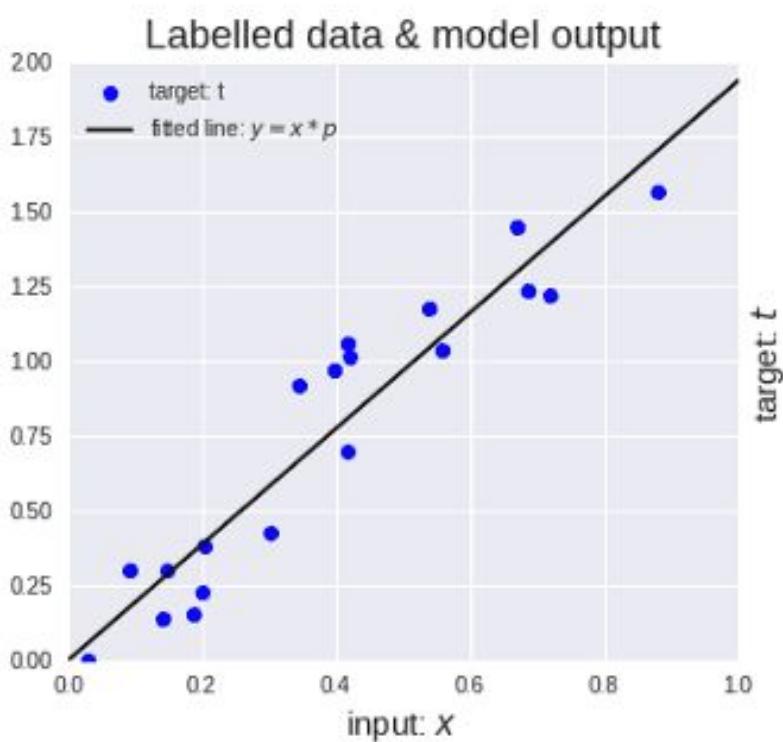
We define a function  $E(W)$  which quantify the error we observe on the examples.

i.e: Euclidean distance between the target and  $f(X)$

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



# 3d visualization of multiple regression



# Linear regression : gradient descent

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

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**.

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

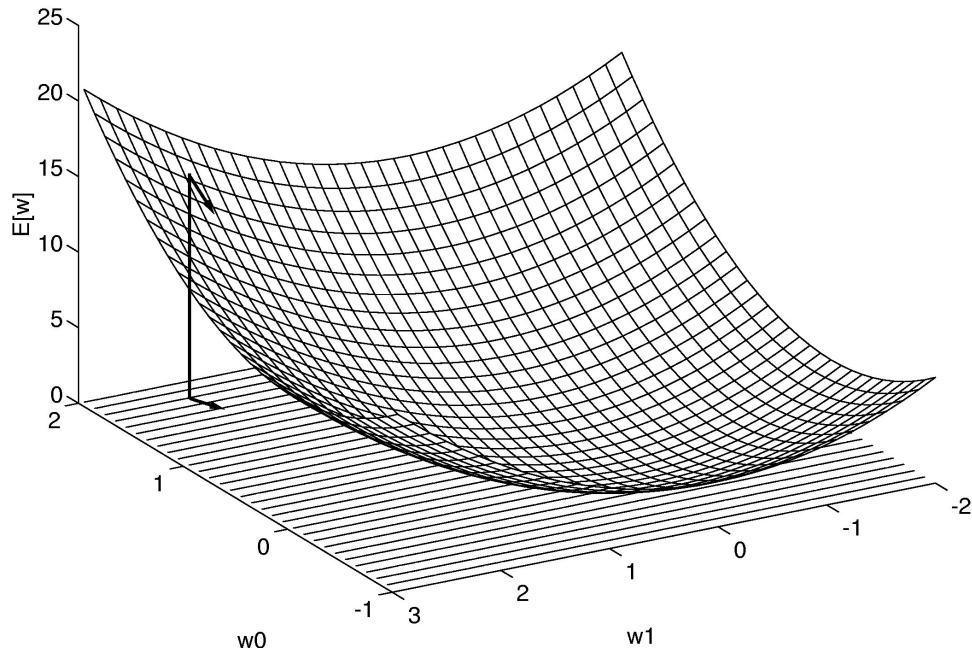
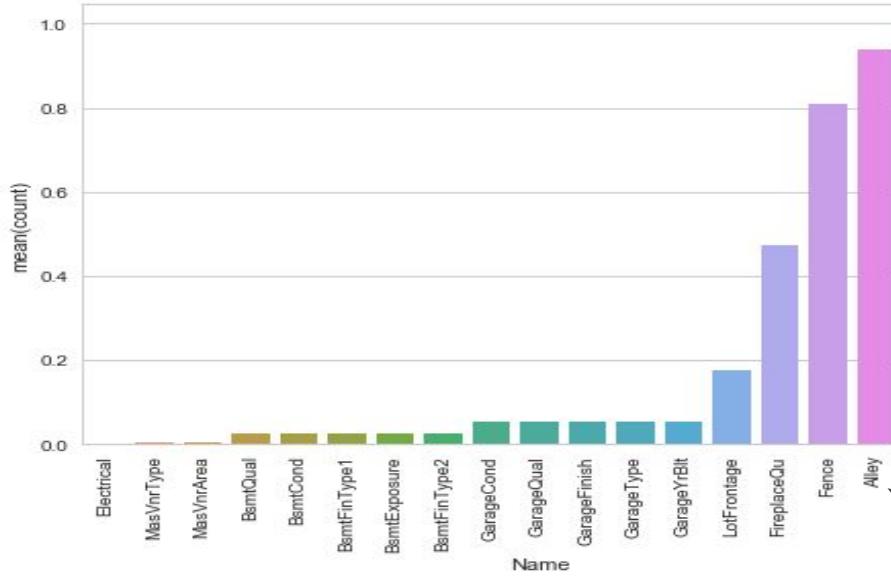


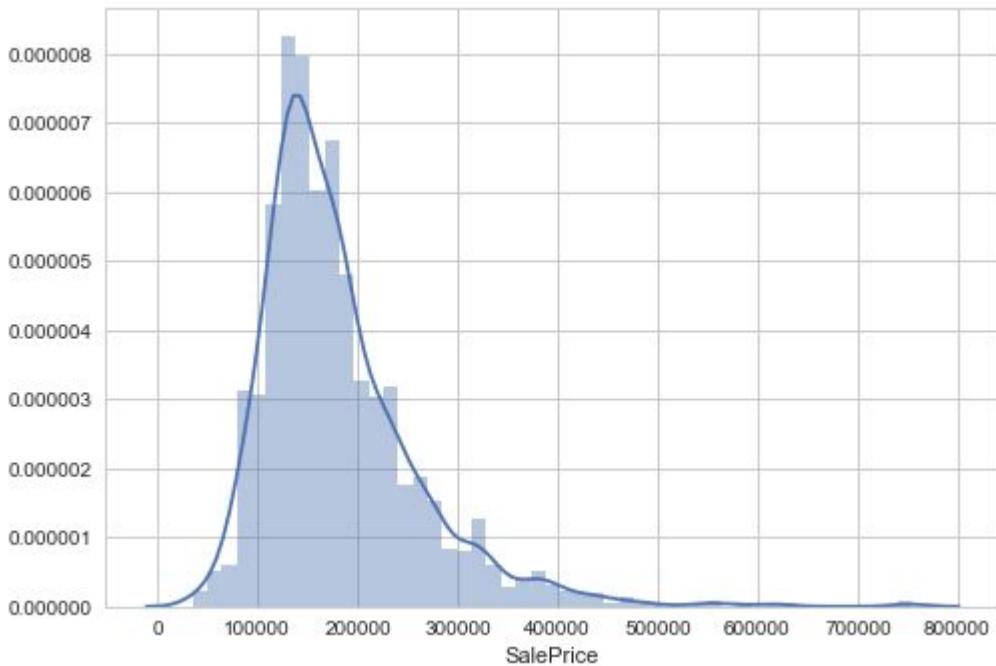
Illustration: Tom Mitchell, McGraw-Hill

# Univariate Analysis - Plot the missing values count

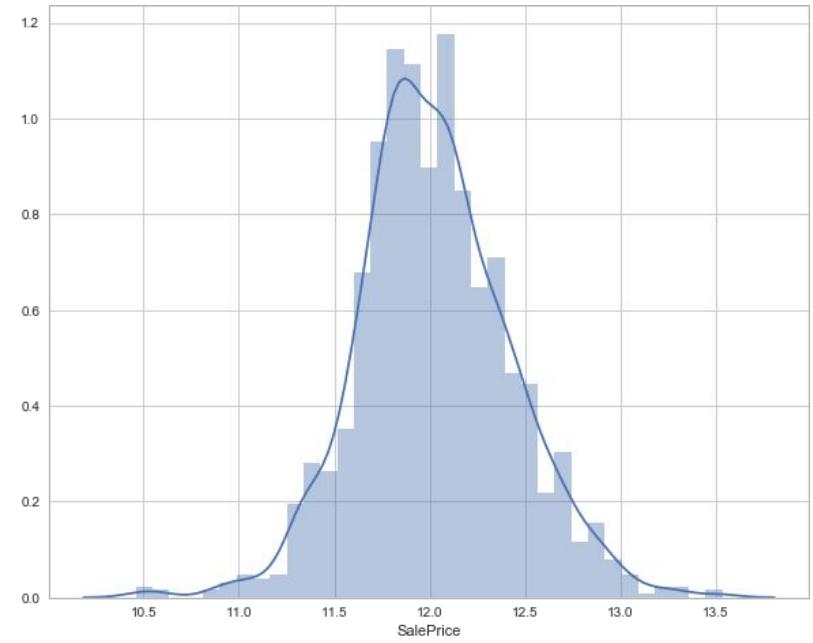


<b>Id</b>	<b>MSSubClass</b>	<b>MSZoning</b>	<b>LotFrontage</b>	<b>LotArea</b>	<b>Street</b>	<b>Alley</b>	<b>LotShape</b>	<b>LandContour</b>	<b>Utilities</b>
1	60	RL	65.0	8450	Pave	NaN	Reg	Lvl	AllPub
2	20	RL	80.0	9600	Pave	NaN	Reg	Lvl	AllPub
3	60	RL	68.0	11250	Pave	NaN	IR1	Lvl	AllPub
4	70	RL	60.0	9550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

# Univariate Analysis - Transformation of Target Variable - Sales Prices

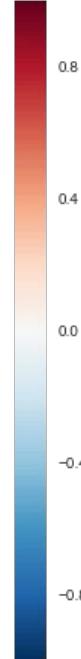
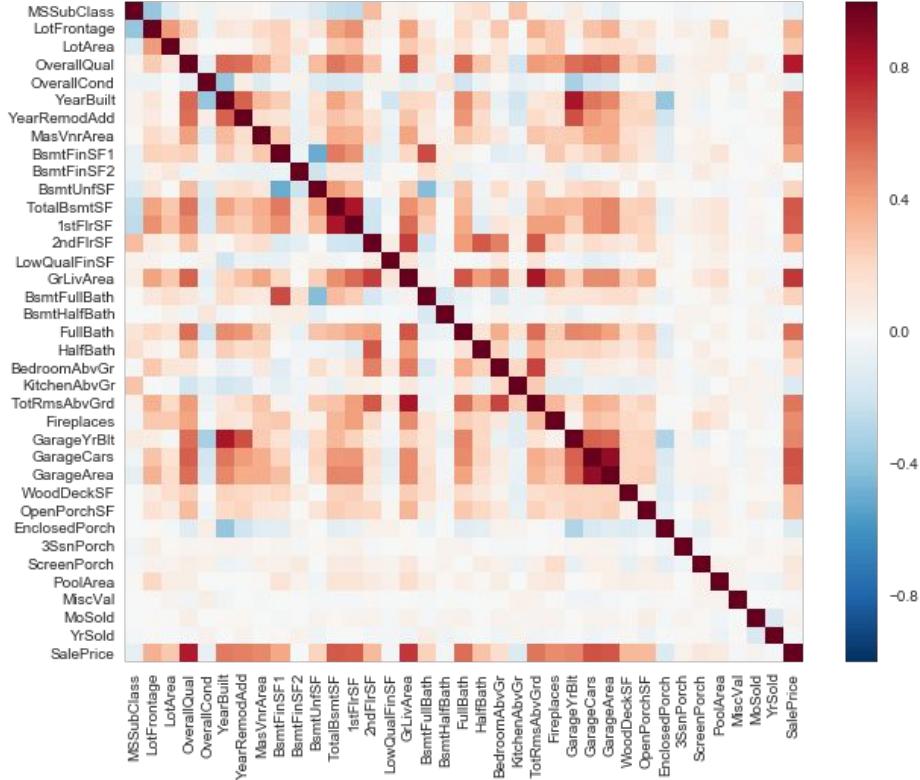


Right Skewed distribution

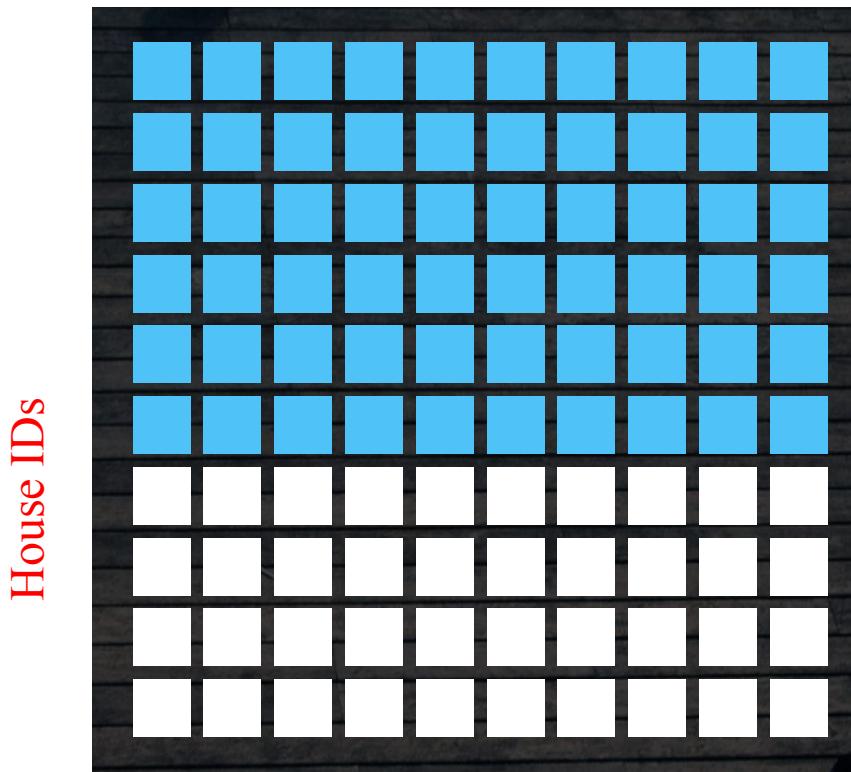


Normal Distribution

# Bivariate Analysis - Transformation of Target Variable - Sales Prices



## Features



## Training Set

# Model Building using training set

## Cross Validation

## Test Set

## Prediction using test set

# Acknowledgement



Bengis Center for  
Entrepreneurship & Innovation

Guildford Glazer Faculty of Business and Management  
Ben-Gurion University of the Negev

## Course Creation and people responsible

Github page, python scripts for week 1 - Minesh Jethva

Administration - Moran Sharon

Ice breaking game, statistical help - Ruth Hashkes

Week 1 presentation - Rahul Veettil

Thank you!

