Gradient Descendants – Ames, Iowa House Data Project

* + Understand the problem
  + Hypothesis Generation
  + Get Data
  + Data Exploration
  + Data Pre-Processing
  + Feature Engineering
  + Model Training – Linear, Random Forrest, Logistic Regression, Lasso, etc
  + Model Evaluation

1. **Understand the problem**

* Project aims at predicting house prices (residential) in Ames, Iowa, USA based on data set provided by Kaggle between 2006 and 2010

1. **Hypothesis Generation**

* What factors do we believe to influence house prices?
* Ho/Ha – using Anova, p-test,

1. **Get Data**

* Data consists of 81 explanatory variables, SalePrice is the target variable

## 4. Data Exploration

Data Exploration is the key to getting insights from data. This is a preliminary evaluation of the data (aka EDA) that helps to summarize main characteristics and generate insights beyond the formal modeling or hypothesis testing task

A good data exploration strategy comprises the following:

1. **Univariate Analysis** - It is used to visualize one variable in one plot. Examples: histogram, density plot, etc.
2. **Bivariate Analysis** - It is used to visualize two variables (x and y axis) in one plot. Examples: bar chart, line chart, area chart, etc.
3. **Multivariate Analysis** - As the name suggests, it is used to visualize more than two variables at once. Examples: stacked bar chart, dodged bar chart, etc.
4. **Cross Tables** -They are used to compare the behavior of two categorical variables (used in pivot tables as well).

## 5. Data Pre-Processing

In this stage, we'll deal with outlier values, encode variables, impute missing values, and take every possible initiative which can remove inconsistencies from the data set.

## 6. Feature Engineering

The goal of feature engineering is to create new features which can help make predictions better.  Opportunity to implement domain knowledge and creativity.

* Geo location (proximities to best schools, commercial centers, transportation stations)
* The S&P CoreLogic Case-Shiller Home Price Indices (West-Central Composite, includes Iowa) – caveat emptor – past performance is not an indication of future performance.
* Fine tune data set

## 7. Model Training and Evaluation

* Linear, Random Forrest, Logistic Regression
* Cross-validation process, and assessment of error (RMSE)

8. **Outcomes and Analysis**