Statistical Modeling using R

# Home Price Prediction Report by Wanogho Darlington

## Overview

* Provided with a dataset containing features and observations of different houses sold in the City of Ames, Iowa. I’ve been tasked to build a statistical model to predict home prices in the area based on several features.
* This report is aimed at going through my analytical process which followed the data scientist workflow of cleaning my data to the final regression/ model building practice to fully develop the results that would be presented in this report.
* Housing is a basic foundation of any society and understanding the intricacies that goes into building a house usually is a highly determinant for estimating the natural mortgage prices before it’s sold to anyone who plans on acquiring a housing property.

## Insights

Performing EDA on the data using PowerBI, the following insights were derived or gotten from the visualizations;

* By average sale prices, Houses of the MSSubClass 60 was discovered to be the most expensive
* Houses located at Northridge (NoRidge) were discovered to be more costly than others.
* There are a total of 5 Zoning Classifications with FV having the highest mean averages while RV/RL had the better mean overall quality.
* In varying the data by grouping the houses by the Lot Area that they occupy, Low Density Residential Zones have larger Lot Areas in respect to the other zoning classifications.

## Methodology

Implementation of the Data Analysis workflow was involved in the carrying out of this project. From processes that involved understanding the data source and its relevance to the project up to analyzing and getting appropriate results.

The Data Analysis workflow implemented is visualized in the flowchart below;

## Results

From the predicted results with respect to the training model;

* Average sale price for the entirety of the data was around 180000USD in comparison to 175000USD of the training data.
* From the predicted results, there were variations in the relevance of the Lot Area in Sale Prices
* Ground Living Area, Neighborhood and Overall Quality of the Housing played a key role in the datasets as the results of the prediction model showed similarities with the trained model.
* In comparing the Average sale price with respect to the Neighborhood of the predicted model to the trained model the top three neighborhoods with the highest overall quality was still consistent with Northridge, Northridge Heights and Stone Brook
* Most Popular type of zoning classification of housing still remains the Low Density Residential house type.

## Recommendations

* It’s ideal to build houses of the MSSubclass 60 as they are the houses that come with the highest quality.