Predicting Home Values Through Random Forest

BY: ALFREDO MARTINEZ

2017

Introduction

Goal

- Predict Home Prices

Client

- Keller Williams Realty, Inc.
- Other Real Estate Investment Companies

Value

- Sharper report analyses to help the negotiation of sales, purchases and other strategic agreements related to real estate properties.

Data

Housing Dataset

- Originally compile by Professor Dean De Cock at Truman State University
- Consists of 2930 observations and 80 explanatory variables for the sale of homes between years 2006-2010.
- Data is from records of city of Ames, Iowa

Weather Dataset

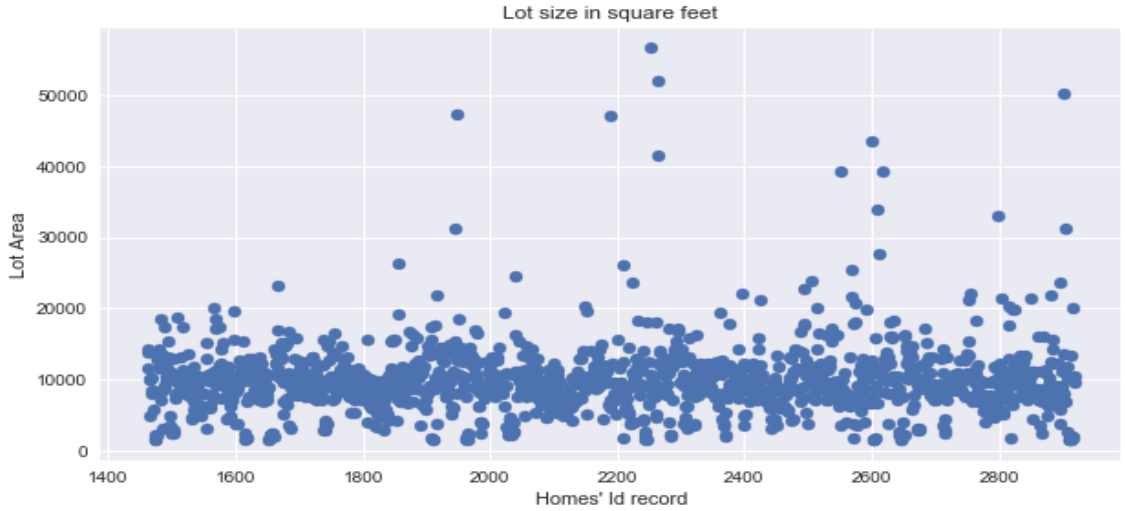
- Subtracted from the National Center for Environmental Information Webpage.
- Includes daily, monthly and yearly records from different weather stations in the city.
- Data is from records of city of Ames, Iowa

Cleaning Data

Bad Data and Outliers

- Random Forest not very sensitive to outliers

Exploratory Data analysis for error type data



Cleaning Data Continue..

Housing Data

- Attached homes' Feature data set to price data set
- Filtered cells to dates 2006 and 2009 to match weather dates available
- Built and sorted a Date Time Index for the sale of homes
- Rest of cleaning was done in future steps after splitting the data to build model

Weather Data

- Subtracted the best dates and weather station available from the data
- Filtered to the most significant weather variables
- Dates were filtered to match housing data sets
- Corrupted values were deleted, and averages obtained

At the end the two datasets were joined

Cleaning Data Continue..

Splitting the data into train and test to fit model

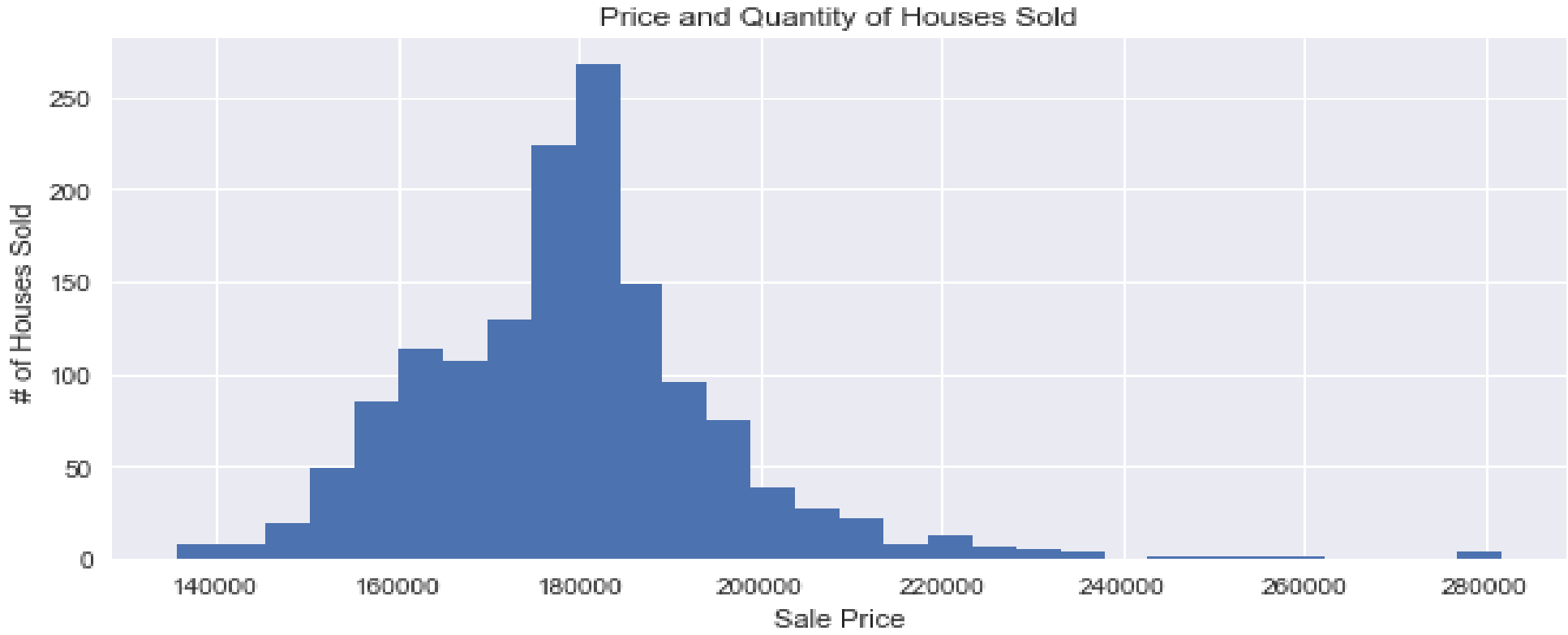
Train

- Filled NaN values with means and modes
- Built dummy columns for categorical variables

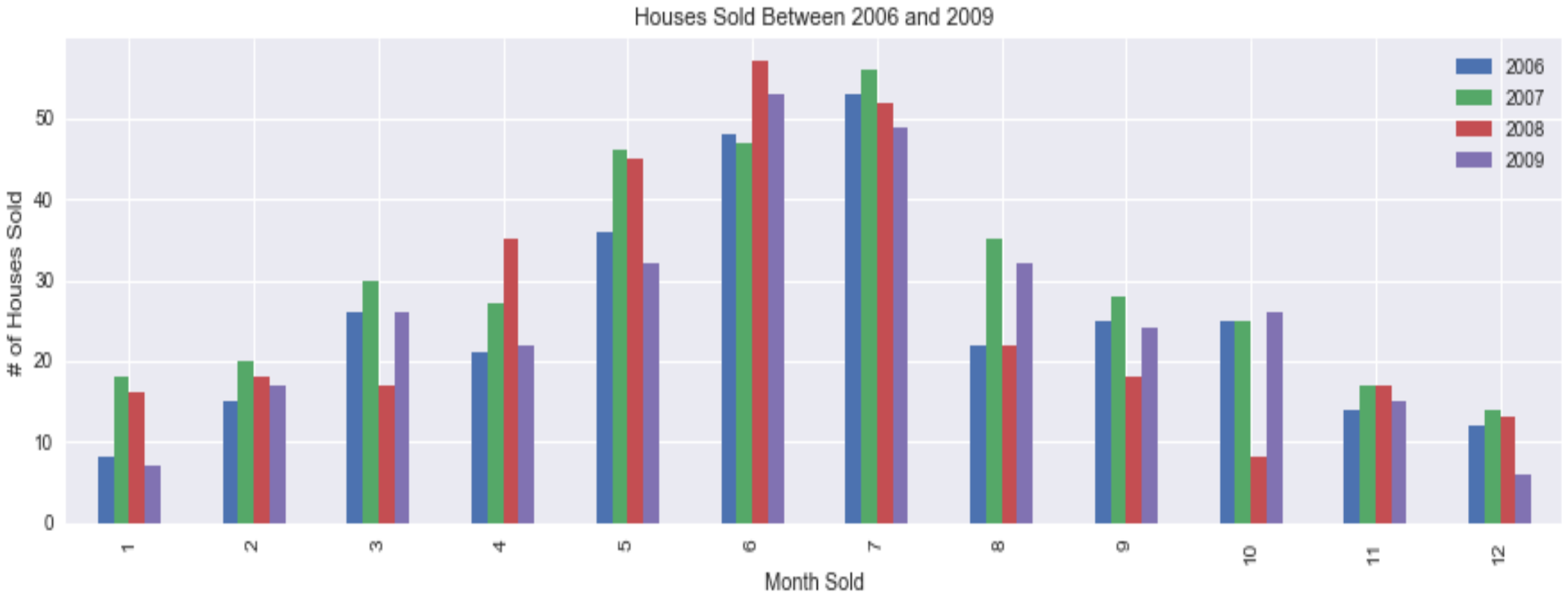
Test

- Used means and mode values from training data for any NaN values.
- Re-index and transfer the dummy variables from out training data

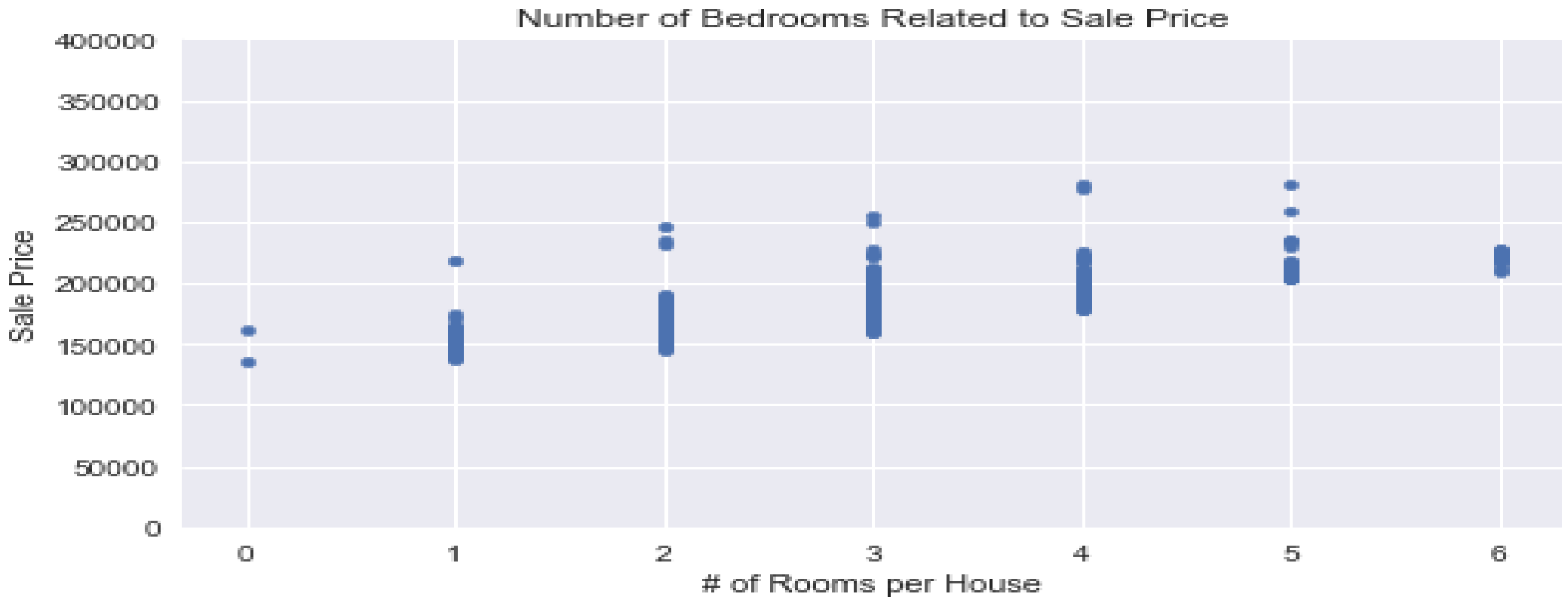
Exploratory Data Analysis



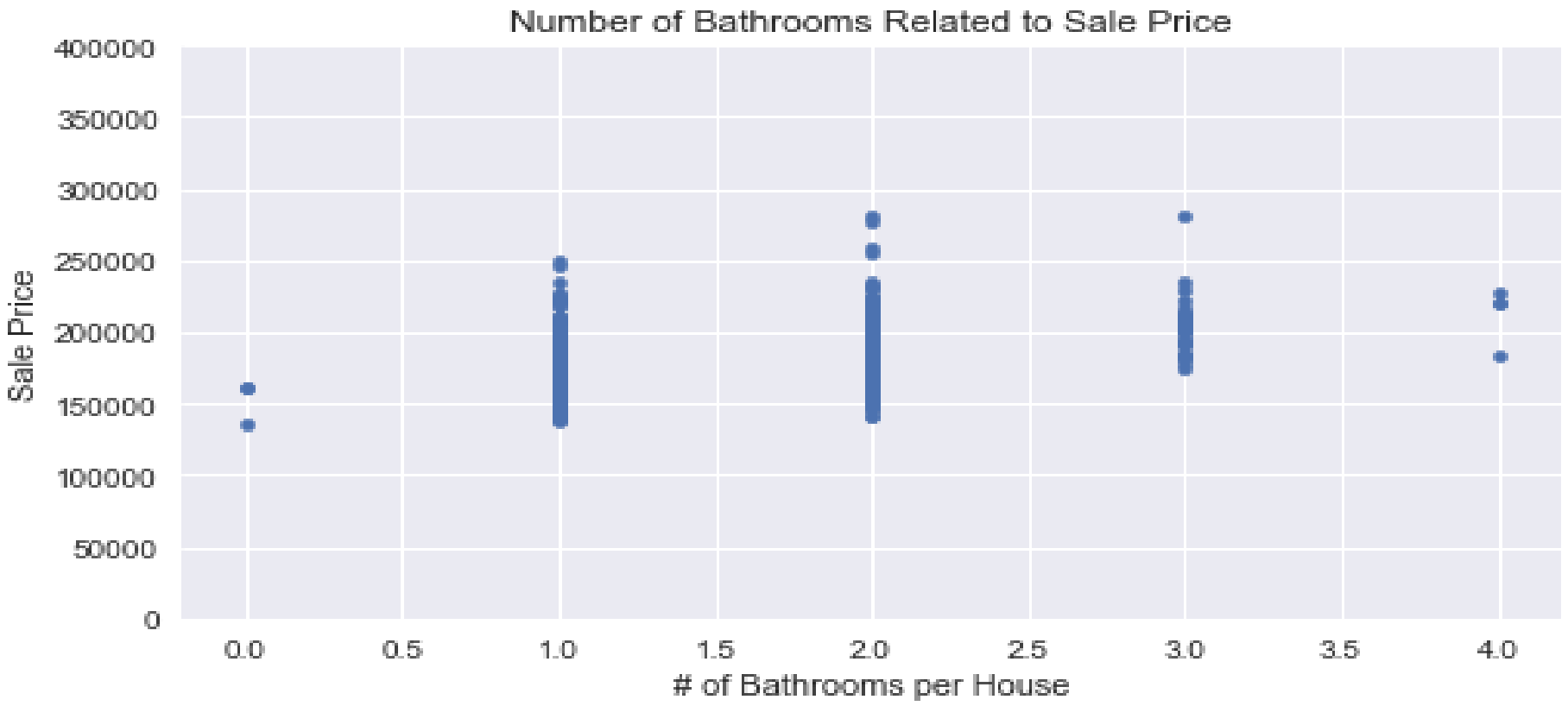
Exploratory Data Analysis



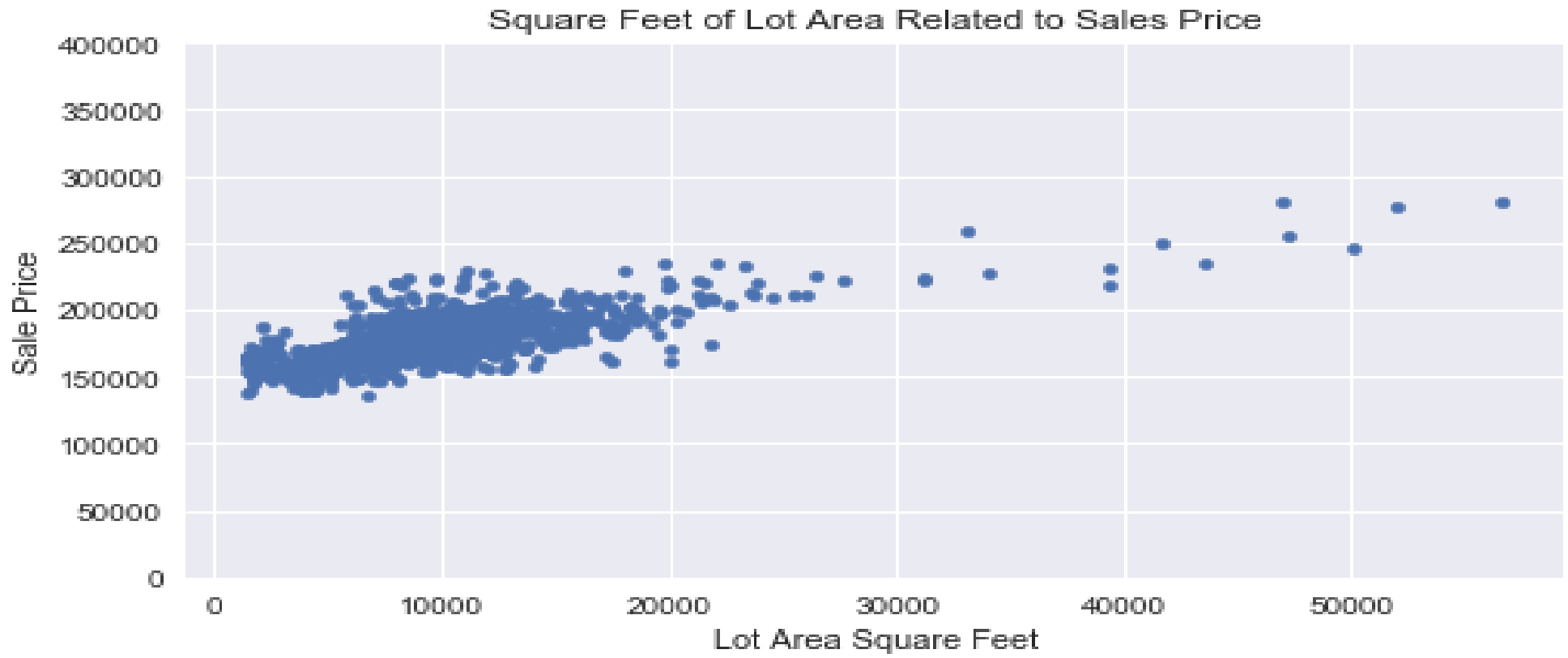
Exploratory Data Analysis



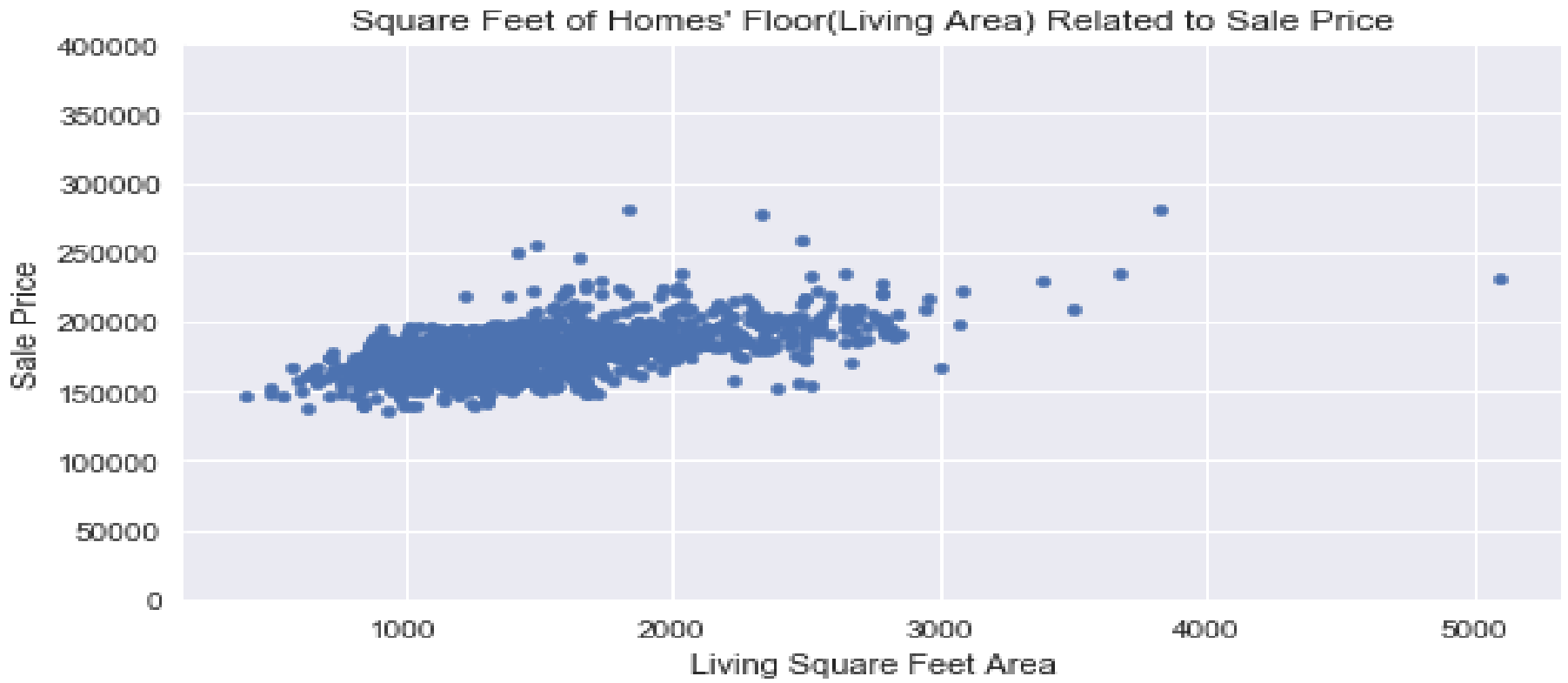
Exploratory Data Analysis



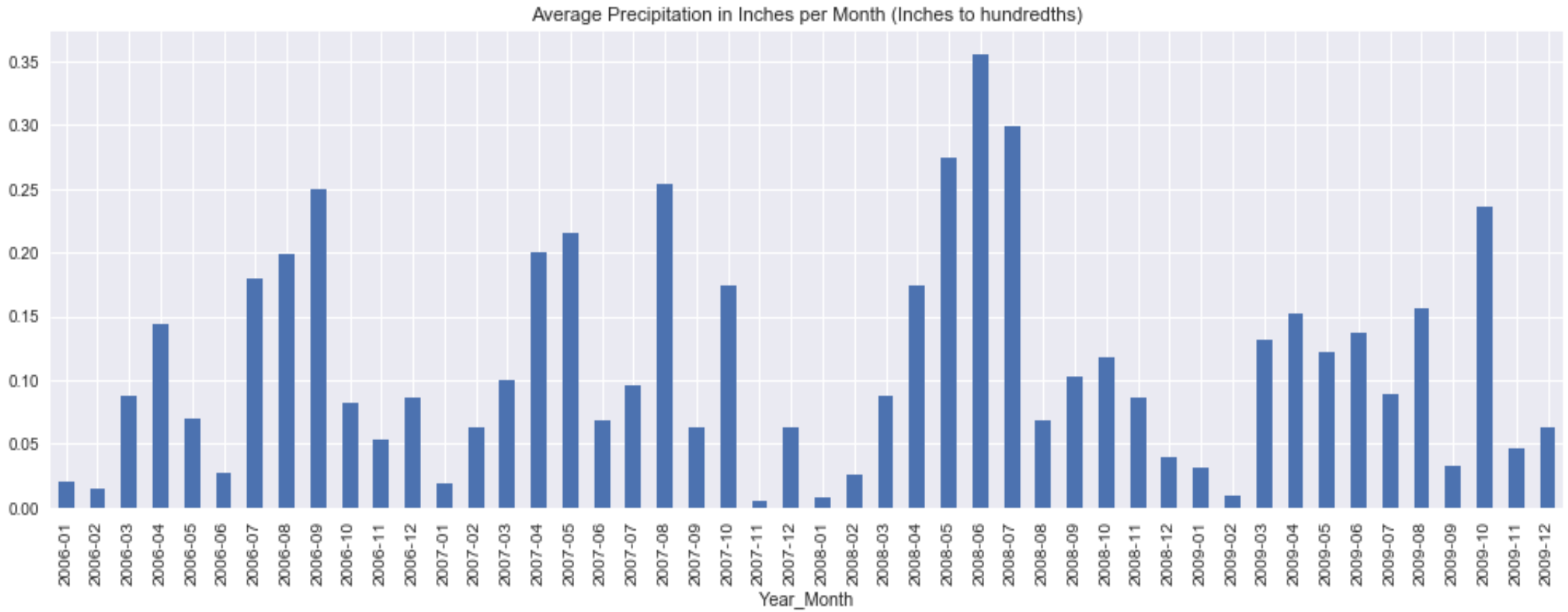
Exploratory Data Analysis



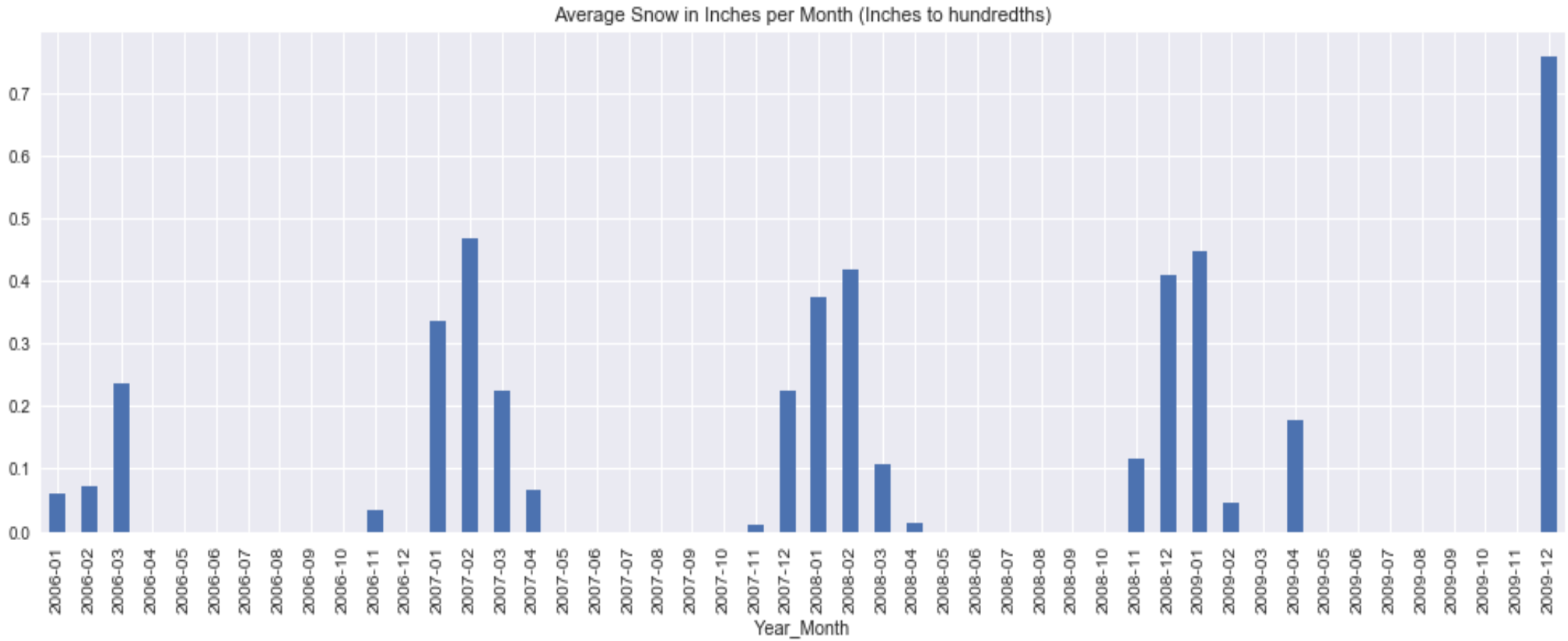
Exploratory Data Analysis



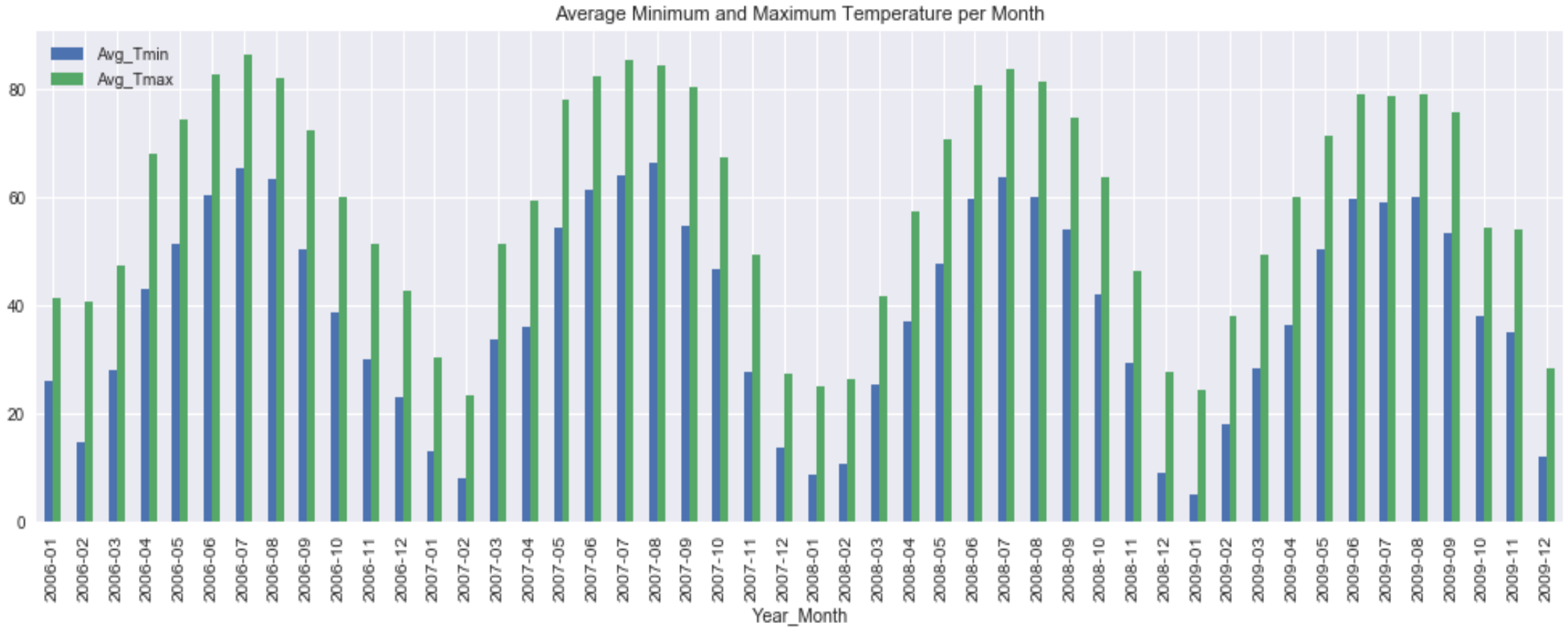
Exploratory Data Analysis



Exploratory Data Analysis



Exploratory Data Analysis



Inferential Analysis

Alpha set at .05

Significant price difference analysis about population.

Central unit vs not a central unit

P-value: 0.0100906212171866 81

One story houses vs Two story houses

P-value: 3.5535259636604621e-13

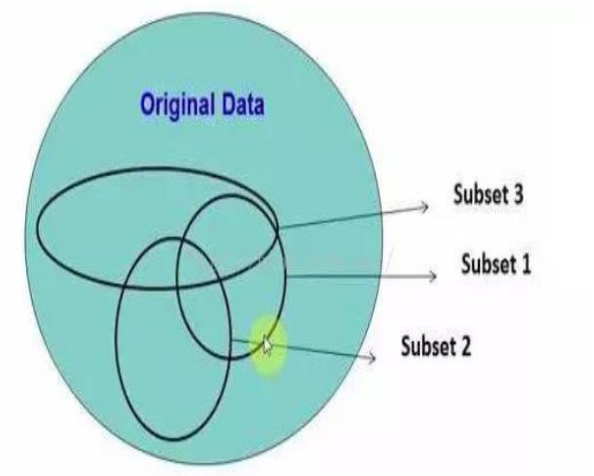
Good privacy fence vs minimum privacy fence

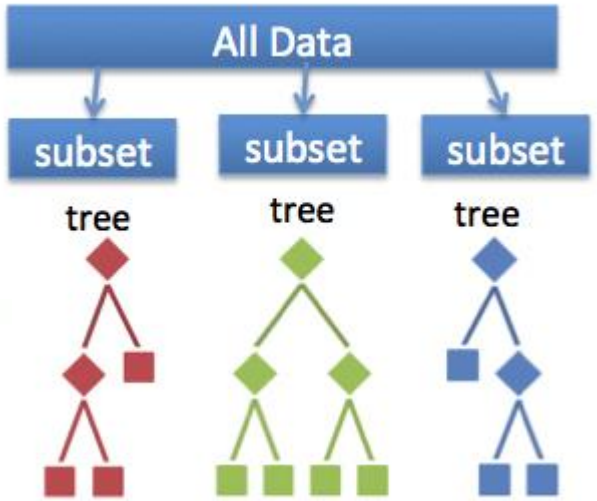
P-value: 0.3700349495497005 1

Modeling and Analysis

Random Forest Predictive Algorithm

- Continues Target variable(Regression)
- Based on Decision Trees Models
- Bootstrapping
- Root Mean Square Error Accuracy
- Feature Importance





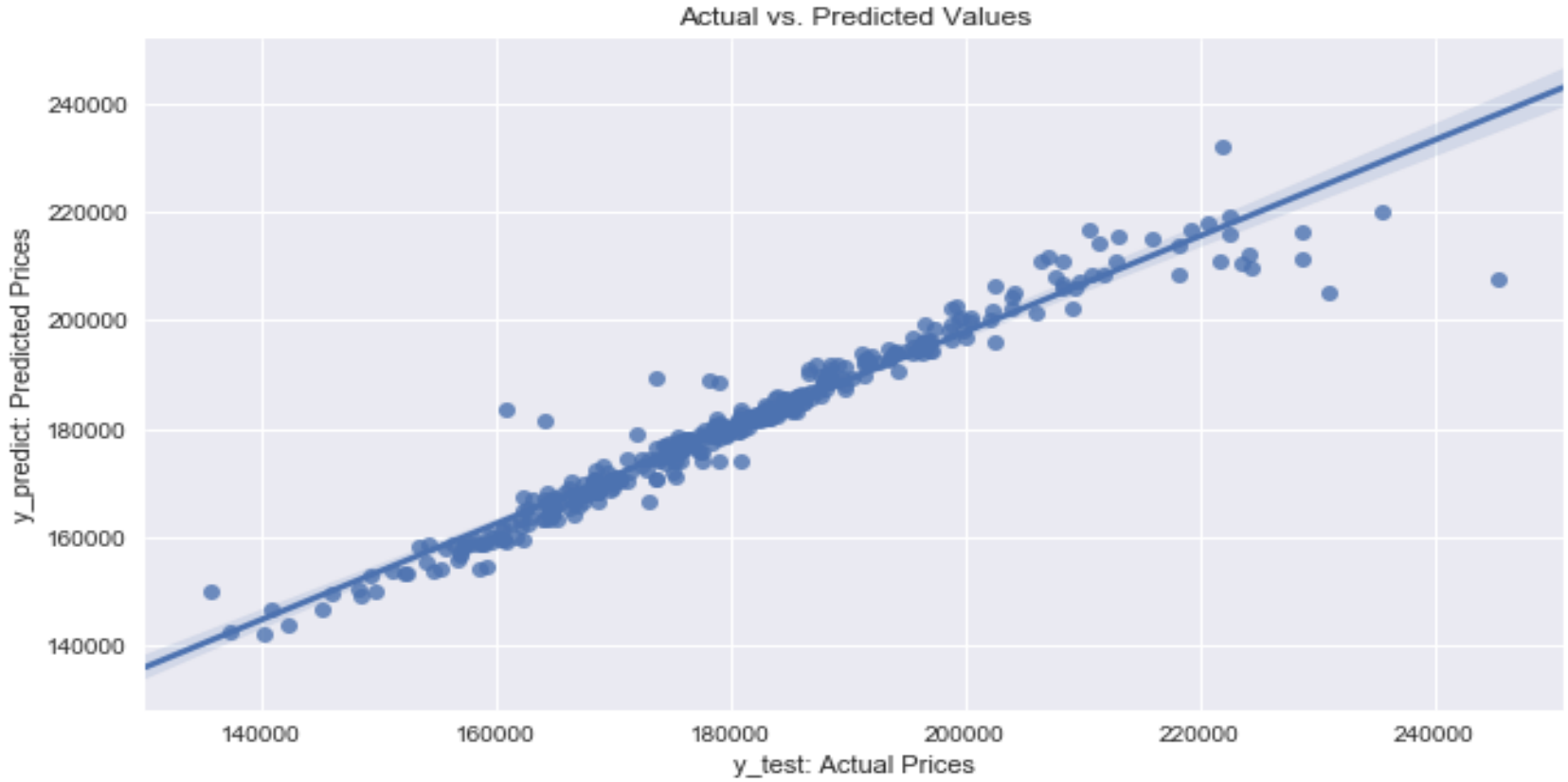
Modeling and Analysis

Main packages and libraries

Scikit-learn and math

- RandomForestRegressor()
- Param_grid(GridSearchCV)
- Fit()
- Best_estimators()
- Predict()
- RMSE = sqrt(mean_squared_error(y_test, y_predict)

Modeling and Analysis

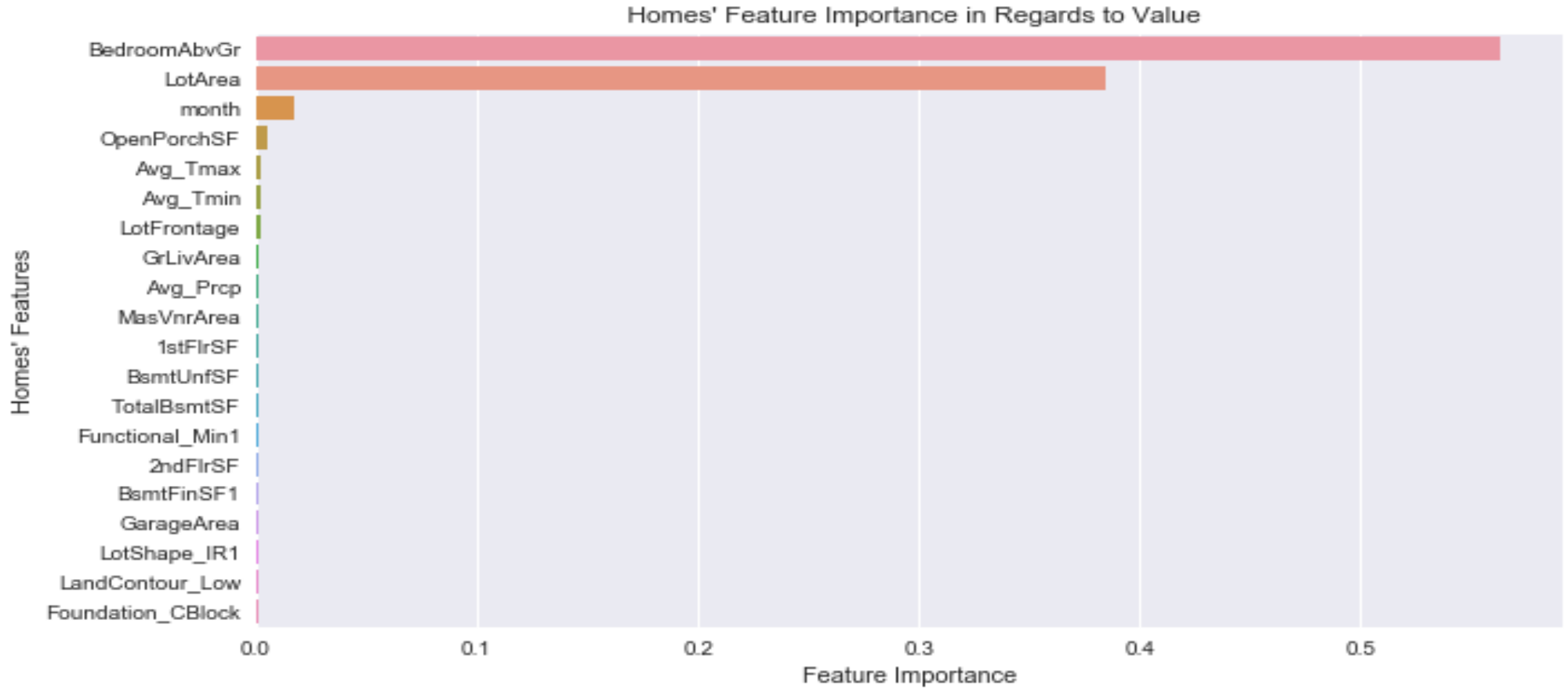


Modeling and Analysis

Feature Importance related to predicted price value

	Features	Importance
20	BedroomAbvGr	5.67e-01
2	LotArea	3.81e-01
34	month	1.72e-02
38	Avg_Tmax	2.60e-03
28	OpenPorchSF	2.22e-03
39	Avg_Tmin	1.96e-03
1	LotFrontage	1.77e-03
12	1stFlrSF	1.55e-03
215	Functional_Min1	1.42e-03
36	Avg_Prcp	1.35e-03

Modeling and Analysis



Thank you!

Any Questions?