Proposal: Getting real with real estate

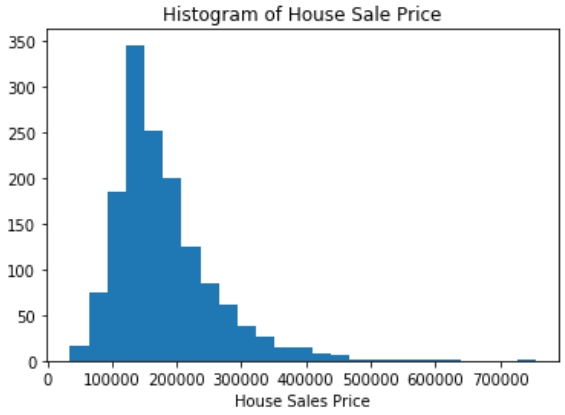
Section 4 Team 2 - Patrick Barranger, Mark Paluta, Subha Paravastu

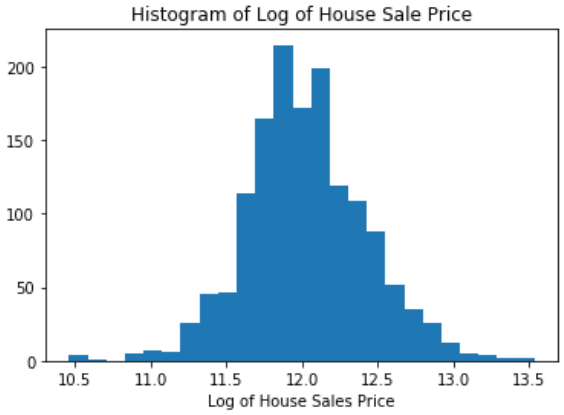
Our project leverages house pricing data provided by Kaggle for a machine learning competition. While we will not be building a ML model (yet), we want to perform an EDA on the data as if we were next going to model. At the highest level, our research will try to answer the question 'What features help best predict final sales price?' Our paper will be structured into five sections: Exploring the data, Handling missing values, Engineering features, Identifying potentially valuable relationships, and Confirming trends (via other house pricing data sets). While our private exploration of the data is not as linear as described, the proposed structure should increase readability and understandability. Below are the five sections, some data we have already uncovered, and our goal for each section.

Exploring the data

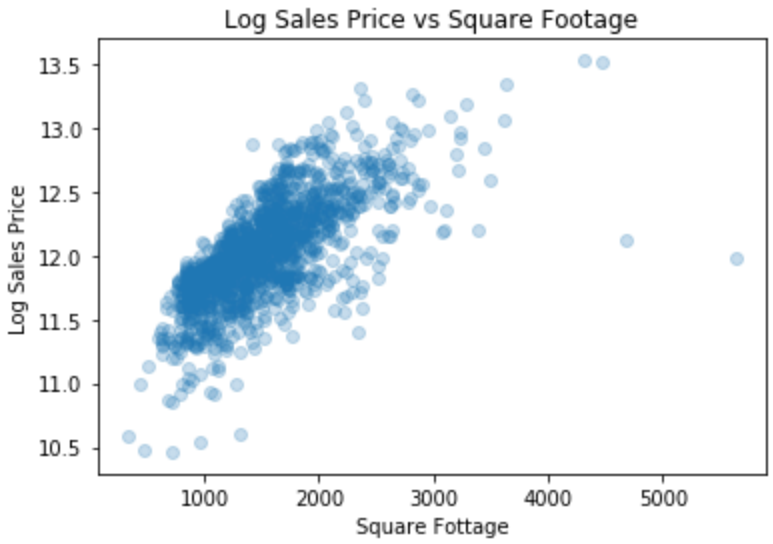
This section is intended to capture the first interactions we have with the data. The issues uncovered here will be worked on further, later in the paper.

Plotting final sales price shows a right skew. To correct the skew we can take the log.





Inspecting the relationship between square footage and price shows that we have 4+ outliers that will require further inspection.



Handling Missing Values

PoolQC	1453
MiscFeature	1406
Alley	1369
Fence	1179
FireplaceQu	690
LotFrontage	259
GarageYrBlt	81
GarageType	81
GarageFinish	81
GarageQual	81
GarageCond	81
BsmtFinType2	38
BsmtExposure	38
BsmtFinType1	37
BsmtCond	37
BsmtQual	37
MasVnrArea	8
MasVnrType	8
Electrical	1

A quick check of the data set shows that we have missing values in many of the columns. To the left is a chart capturing the column and the number of null values it contains.

Many of these nulls make sense, like Garage*, Alley, and Fence. However, there are some nulls that seem to be missing at random (MasVnrArea, LotFrontage, etc.). We will need to determine if we impute these values, delete them, or just leave them.

Engineering Features

We have continuous, discrete, categorical and ordinal data in our data set. We will need to be thoughtful about how we handle each data type. Additionally, we might want to consider what transformations would be needed for a model to correctly handle the categorical and ordinal data.

```
object     43
int64      35
float64     4
```

Identify potentially valuable relationships

Once the data has be explored, cleaned, and had the missing values handled we will explore the relationships between the features and Sales Price. Some of the questions we hope to answer include:

- What features have the strongest correlation with sales price?
- Does average sales price vary by neighborhood?
- Do different zonings lead to different key features? (Ag vs residential)
- Do building materials have an impact on price?
- What is the value of adding features (garage, pool, etc.)?
- What is the impact on the age of the house and final sales price?
- In which year the houses have been sold at the highest price? Did housing bubble, credit crisis and great recession of 2007-2008 have an impact on Sale Price?

Confirm Trends

We have identified two additional data sets (North Carolina and California) to combine with our Iowa dataset. Between the three datasets we have the following overlapping columns between at least 2 data sets: price, bedrooms, bathrooms, house size, plot size, sales status, and year built. Both additional datasets will need to be cleaned before combining with the Iowa data set. (For example, plot size is in acres and will need to be converted to square feet.) Additionally we may need to index house pricing since since identical homes in different markets can sell for drastically different prices. Finally we will need to smartly deal with null values for data sets where columns do not overlap. Our ultimate goal with this cleaned and combined data is to investigate if the trends identified in the Iowa dataset persist in housing markets elsewhere in the United States.

Links:
Iowa House Pricing Dataset -
https://www.kaggle.com/c/house-prices-advanced-regression-techniques
North Carolina Dataset -
https://github.com/MIDS-INFO-W18/house_price_sec4_team2/blob/master/woodard.xls
California Dataset -
https://github.com/MIDS-INFO-W18/house_price_sec4_team2/blob/master/RealEstate.csv