# Model: Estimating Housing Prices Across the US

Kenneth Truong

# Background

- New graduates from college tend to look for jobs in very aggressive and popular job markets
- The housing prices at these job markets tend to be more than they can afford.

# Deliverables:

- Predict bedroom prices using income, population size, and state data using Zillow Data.
- Provide list/data of least affordable/most affordable places in the US.

# Beneficiaries:

- Job seekers looking for job markets/areas that are more affordable
- Real estate developers looking to develop housing in up and coming areas.
- Congressmen who want to pass legislation to make housing more affordable.

# Data Wrangling

- Load datasets and Create DataFrames
  - Used pd.read_csv to load 4 datasets
- Check DataFrames

- Select necessary columns
  - Retained 2015 and 2019 data in bedroom sets and all of income set

# Data Wrangling(cont.)

- Melt DataFrames
  - Used pd.melt to have all of 2019 and 2015 months under one column.
  - Grouped all of them and took the mean to get one yearly value.
- Merge DataFrames
  - Merged 2015 and 2019 versions of datasets together.
  - Then all of the bedroom sets and income sets.
- Treat NaN values
  - Using OLS function, get coefficients from 2019 set to fill 2015 NaN sets.
  - Then use same OLS function on different bedroom sets to fill non-corresponding values.

# Data Wrangling(cont.)

- Feature engineering:
    - Created new column which coded the States numerically to compare them.
    - Also created Price/Income Ratios of different bedroom rent prices.

# Least Affordable/Most Affordable Counties

## Least Affordable:

1. San Francisco County, California
   0.517221
2. Queens County, New York
   0.472453
3. Suffolk County, Massachusetts
   0.458542
4. Miami-Dade County, Florida
   0.444621
5. Orleans Parish, Louisiana
   0.431616

## Most Affordable:

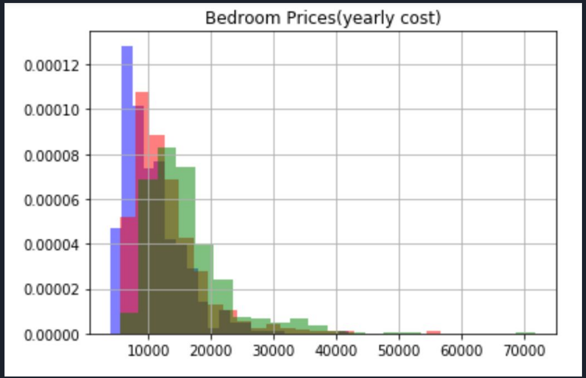
1. Johnson County, Missouri
   0.093180
2. Allen County, Indiana
   0.090751
3. Coryell County, Texas
   0.090472
4. Hardin County, Kentucky
   0.088275
5. Cole County, Missouri
   0.076014

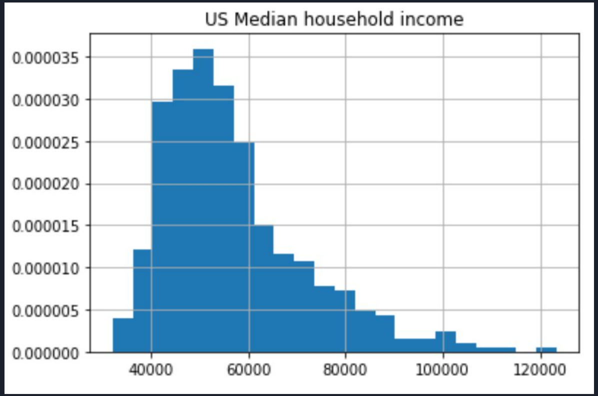
# Checking collinearity and correlation between variables

- OLS function provided R^2 values between different variables
- Also used the pearsonr function to hypothesis test for correlation
- Conclusion: While income and population were correlated with bedroom price, the R^2 were not high.

# Checking Normality of Data:

- Plotted histograms of the data
- Tested normality using Chi-Square Tests.
- Conclusion: Data is not normally distributed.



# Models Used

- Linear Regression
  - Predicts linearly correlated variables together
- Random Forest Regressor
  - More robust
  - More accurate

# Linear Regression

- Split into training and testing sets (80%/20%)
- Trained on the training set and tested on the 20%
- Result: $R^2$: 0.261

# Random Forest Regressor

- Split into training and testing set (80%/20%)
- Trained with GridSearchCV with the following Hyperparameters:
  - Cross-validation: 3
  - Max_depth: list(range(1,20))
  - N_estimators: list(range(1,20))
  - Max_features: list(range(1,3))
- Tested on 20% data
- Result: R^2: 0.535