**Dataset**

The Zillow Rent Index is the median estimated monthly rental price for a city, reported in US dollars. It includes multifamily, single family, condominium, and cooperative homes in Zillow’s database, regardless of whether they are currently listed for rent.

The dataset is comprised of 13,131 observations (rows) and 81 variables (columns). The variables consist of:

* City Code – A unique identifier for each city in the dataset
* City
* Metro – Metropolitan area
* County
* State
* Population Rank – Cities are ranked 1 through 13,131 with 1 being the most populated and 13,131 being the least.
* The remaining 75 columns represent months - November 2010 through January 2017 and contain the median estimated monthly rental price for each city.

**Business Problem**

We are real estate developers, deciding on where we should build our next rental properties. We are primarily interested in building in areas where we expect rents to increase over time. Business questions that we want to address with the aforementioned dataset -

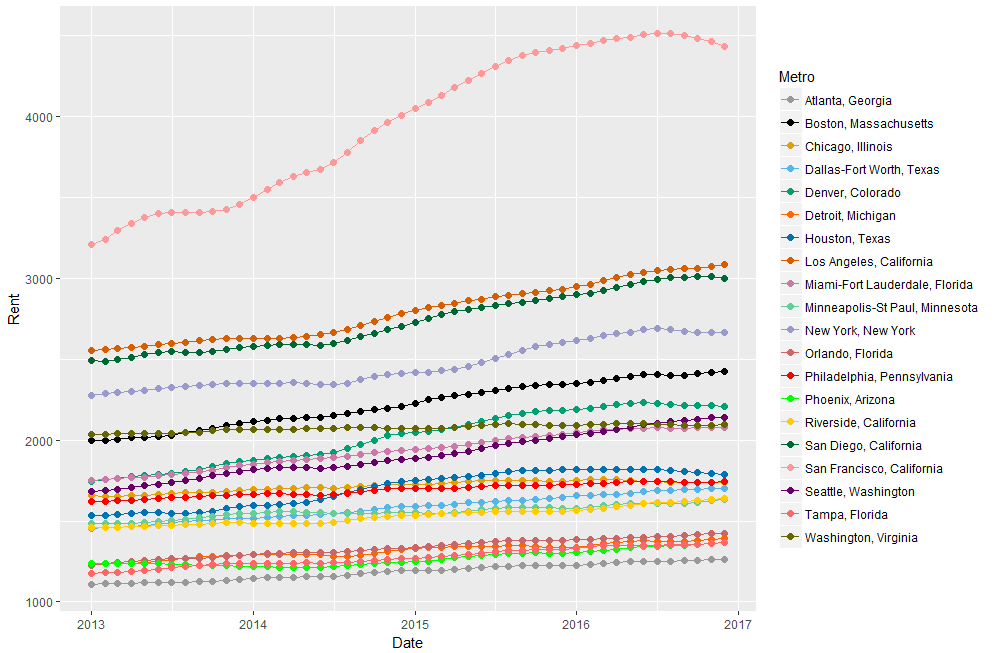
* Given historical monthly rent trends, where are rents likely to increase/decrease?
* What is the expected monthly rent for a given city, one year in the future?
* Given changes in median household income and population, predict where rents are likely to increase?

**Data Cleaning and Pre-Processing**

* Zillow began estimating rent prices at different times for each city, resulting in a significant number of missing values for a large number of cities for November 2010 through March 2012. We decided to just keep the years that have full value for each month, that is 2013-2016. Fortunately, it makes sense since we had several years of data to work with and the missing values were concentrated in the first couple of years. Additionally, we decided that the rent estimates for these years were too far in the past and were likely not a good indication of future rent prices due to changing real estate market conditions. We therefore excluded 2010 through 2012 data from our dataset.
* The dataset is in wide format with each month representing a column. This is a pretty common issue in data cleaning, when column headers are values instead of variable names. In order to appropriately analyze, graph and model the data we reshaped the dataset to long format, creating a Month column and a Price column.
* Useful variables, such as demographic information, are not in the dataset. We believe it's necessary to add census demographic information to improve the accuracy of our rent forecasting. So we downloaded American Community Survey (ACS) 1-year census data of 2013-2016 from <https://factfinder.census.gov> for population and median household income data by county, and added these columns to the dataset.
* Columns unrelated to our business problem (City Code and Population Rank) or needed to be tidied for future analysis purpose, were removed and adjusted. 1 year lagged rent, median household income and population were added to each row.

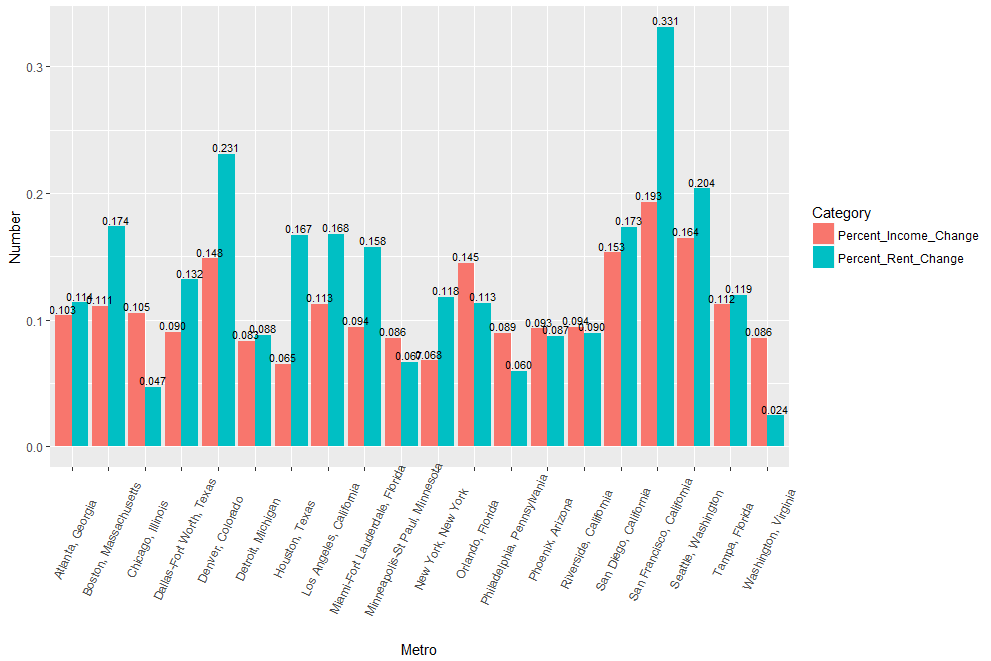
**Exploratory Data Analysis and Visualization**

As part of the initial analysis of our dataset, we analyzed monthly rent increases for the top 20 metropolitan areas to give us a better understanding of the historical data. From the line chart, majority of top metro areas have seen continuous rent increases from 2013 to 2016, with San Francisco achieving the highest dollar gains.

****

We also ran a quick analysis to compare changes in income and rent for the top 20 metropolitan areas from 2013 to 2016. All areas achieved rent and income increases.

Increased median household income could be used as an indicator of an improving economy, which supports rising rental rates as people have more money to invest into housing. As this bar chart shows, 65% of metro areas see higher percentage of rent change than income; Boston, Denver, Houston, Los Angeles, Miami, New York, San Francisco see significant difference (>5%) between changes in rent and income.

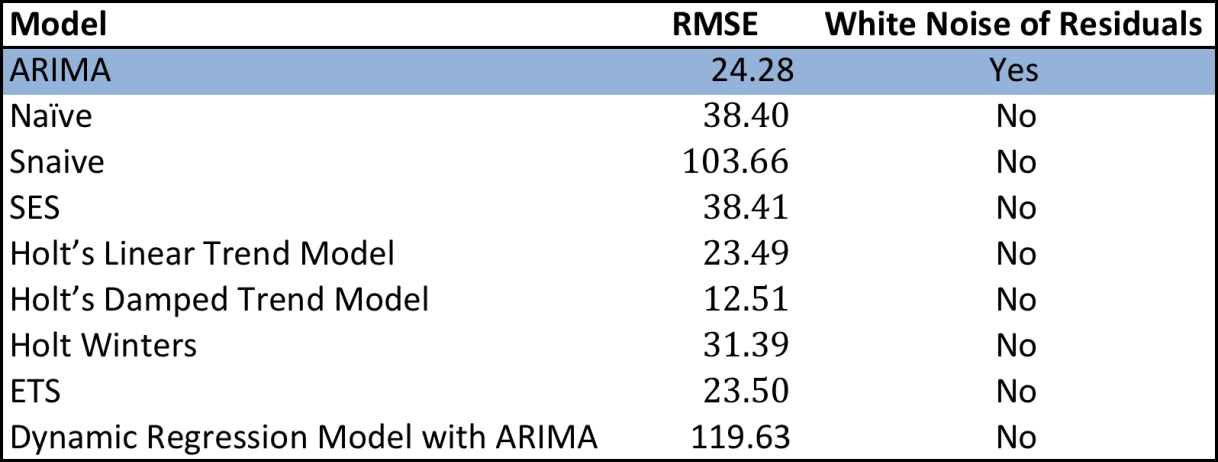


**Rent Forecasting**

**Time Series Methodology**

Our dataset contained four years of monthly rent, 2013 - 2016. We split the data into a training set(2013-2015) and a test set(2016). As our dataset contained numerous metro areas we looked specifically at the New York, NY metro area to train and test the model in order to determine which modeling technique was most effective. We attempted ARIMA model, Naïve, Snaive, SES, Holt’s Linear Trend Model, Holt’s Damped Trend Model, Holt Winters, ETS and Dynamic Regression.

For each modeling technique, we used the training set to build the initial model and the test set to check forecast performance on 2016 data (for our dynamic regression model the dataset also contained metro area population and metro area median household income). Using a training and test set helps to overcome the problem of overfitting. We computed the accuracy of each time series model using the forecast errors (the difference between actual 2016 monthly rent and our forecasted rent in the test set). Ultimately, the ARIMA model was the best time series modeling technique as it had both a low RMSE and white noise residuals. A lack of white noise residuals implies that the forecast method used did not capture all of the available information in the data.



We then ran the ARIMA model on each of the top 20 metro areas then computed a combined RMSE and a combined MAPE, which we compared with the decision tree and random forest models (described below). We found that the ARIMA model produced results with the least errors.

**Decision Tree Methodology**

Time series modeling (excluding dynamic regression) only takes time series data into consideration. In contrast, decision trees are able to incorporate other factors that may influence the outcome of our rent forecast. We therefore used a decision tree modeling technique to analyze our data. We built two trees, one with monthly rent as the target variable(Tree1) and one with monthly rent change as the target variable(Tree2). Both trees used Metro, Year, Date, State, Month, one year lagged rent, one year lagged income and one year lagged population variables for prediction. Both trees were built on the training set (2014 - 2015), pruned to avoid overfitting the data, and then tested on test set (2016).

In Tree1, lag\_rent was the most important variable. Date and State also played some part in splitting the data in the latter part of the tree. In Tree2, Metro was the most important variable but lag\_rent, lag\_income, Date, lag\_population and Month also appeared several times during the splitting procedure. Tree2 is more complicated and has more leafs than Tree1.

MAPE(percentage errors) is scale independent and is often used to compare models with target variables in different units. We therefore compared the MAPE of the two trees and rejected Tree2 as it had a higher MAPE.

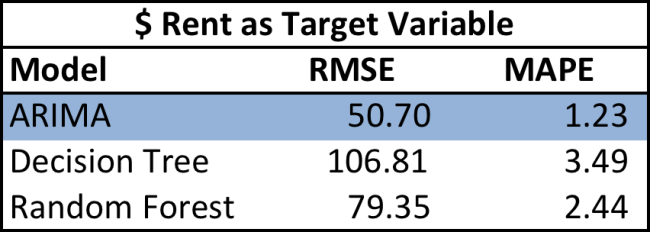
**Random Forest Methodology**

Random forest modeling has the advantage of preventing overfitting and limiting error due to bias and therefore tends to yield useful results. It's more robust than a single decision tree. Due to this, we built two random forest models with the same target variables as our decision trees: one with rent as the target variable (Rent.rf) and one with rent percentage change as the target variable (Change.rf).

As the MAPE for the rent percentage change model is greater than the MAPE for rent as the target variable, we rejected the Change.rf model.

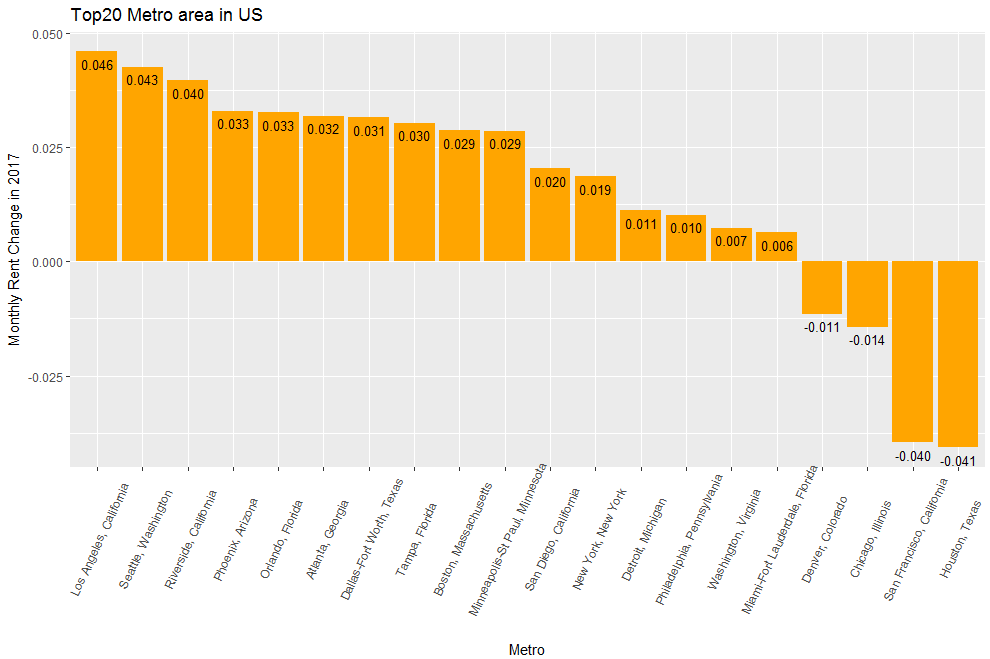
Additionally, the RMSE of the random forest model is less than that of the decision tree model, we therefore rejected the decision tree model.

The RMSE of the random forest model with rent as the target variable is greater than the RMSE for the ARIMA model. Due to this we rejected the random forest model and used the ARIMA modeling technique for our final forecasting of 2017 rental rates.



**2017 Year-Over-Year Forecasted Changes in Rent, Top 20 Metro Areas**

Our research concluded that out of the top 20 metro areas in the US, rental rates are likely to increase the most(4.6%) in Los Angeles, California; followed by Seattle, Washington and Riverside, California. In contrast, rents on average in San Francisco, California and Houston, Texas and are expected to decline 4.0% and 4.1%, respectively.



**Forecasting 2017 Rent for Top 20 Metro Areas Using ARIMA Model**