**Project 1 Summary: Group 7**

**Group Members:**

Ralph Baroud, Daniah Al Bayati, Jeremy Hartnett, Shayon Keating, David Mitchell, and Velkis Moreno

**Problem Statement**

When looking for the right Atlanta property, there are too many variables and details to consider that makes the buying decision that more difficult. For example, real estate sites like Zillow or Realtor.com are not providing an optimized an in-depth analysis to quickly let the buyer/investor know if a specific neighborhood’s listing has been sitting on the market due to other underlying factors nor do they provide future listing forecasts of pricing and number of days to expect they will be on the market.

Although, the average buyer/investor makes decisions based on public information, websites or even gut feeling, through the use of historical data, there is potential to provide forecasts for future listings.

**Motivation**

Understanding the historical trends of Atlanta’s real estate is crucial in determining the future state of the market as well as making informative decisions that can lead to more bang for your buck. Therefore, as a group, we felt that if we can find data trends between a zip code, their list price, and their days on the market, we may present rich opportunities within the real estate market.

Our goal is to provide a potential buyer or investor an understanding of when is the best time to buy, what zip codes have the most days on the market, and if there is any correlation between listing price and days on the market for a specific zip code.

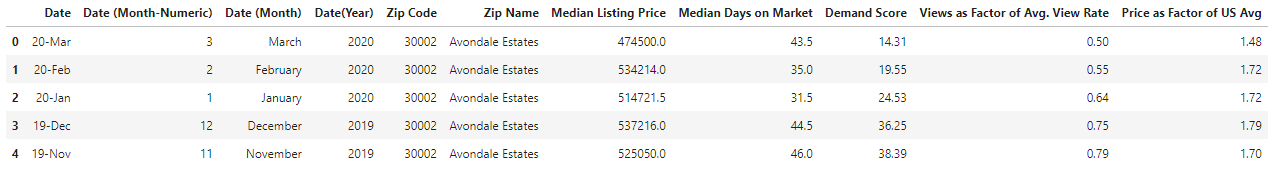
**Hypothesis**

Through the use of historical data, we can describe variation in sales price, time on market, and demand for listings

**Data Set**

Source: <https://www.realtor.com/research/data/>

Sample data:



**Area covered:**

* **41 zip codes** mapped to the metro Atlanta area (mainly inside the perimeter/Fulton County)
* **45 months of data –** from July 2016 to March 2020
* **1,800 data points**

**Data Definitions**

**Median Listing Price:** The median listing price within the specified geography during the specified month.

**Median Days on Market:** The median number of days property listings spend on the market within the specified geography during the specified month. Time spent on the market is defined as the time between the initial listing of a property and either its closing date or the date it is taken off the market

**Demand Score:** The demand score is an index representing a zip code, county or metro’s listing page views per property ranking compared to other zip codes, counties, or metros.

**Views as a factor of Avg. View Rate**: Count of views a typical property receives in the specified geography divided by the count of views a typical property receives in the US overall during the same month.

**Price as Factor of US Avg**: The median listing price within the specified geography divided by the median listing price for the US overall during the same month.

**Analysis & Findings**

1. **Zip Code**
   * There are significant differences in Median Listing price per zip code (Box plot of Median Listing Price per Zip Code)
   * There are significant differences in Median Days on Market per zip code (Box plot of Zip Code vs. Median Days on Market
2. **Year**
   * **Year vs. Median Listing Price**: There is a statistically significant difference in Median Listing Price between years (Anova pval- .001); paired T tests shows significant different years: 16’ vs. all; 17’ vs. 18’.
   * **Year vs. Median time on market:** There is a statistically significant difference in Median Time on Market between years (Anova - 8.94e-05); paired T tests shows significant different years: ‘17 vs. ‘16, ‘19,’ 20; ‘18 vs. ‘19, ‘20
3. **Month**
   * Median Sales price by month – there is not a statistical difference in median listing price by month (Anova, p = .977). Paired T-test does not show significant difference in any year
   * Month vs. Median time on market – there is a significantly different (Anova: p=1.06e-77); paired T-tests show significant differences in practically all of them.
4. **Demand Score**
   * Slight positive correlation between Demand score vs. Median Sales price as shown by linear regression (linear regression: r squared = .0097)
   * Slight negative correlation between Demand score vs. Median time on market (linear regression: r squared = .027)
5. **Time on Market**
   * Slight positive correlation between Time on Market vs. Median Sales Price (linear regression: r squared = .11)

**Discussion/Post Mortem**

Based on our findings, we were able to determine a correlation around pricing, days on market, dates, zip code and demand.

Ex: In 2018 we see higher average prices, with less days on the market, but a higher demand, versus in 2019, where we see lower average pricing, with more days on the market, and less demand.

Using big data analytics to determine trends allows us to forecast future listings even before they hit the market.

**What does this mean for real estate investing?**

1. For a buyer or seller, this allows them to determine which time of the year would be best to buy or list a property.
2. As for the investor, it allows them to get ahead of the market, alleviates the uncertainty behind pricing, and even predict profitability.

**Next Steps**

* **Find a more specific dataset that will allow analysis on a per property basis**, add in more variables unique to each property, e.g. number of bedrooms, neighborhood amenities, etc.
* **Create filterable interactive geo-visualizations** of our data that would allow for simpler processing of the data enable insight generation for potential buyers/investors
* **Build a multi variable linear regression and more advanced machine learning models** to transition from descriptive models to predictive models of the data. This step would allow for price forecasting for a property/neighborhood given a set of filterable features.
* **Build a platform / website for our model** and begin A/B testing given preliminary user feedback