Group 41 Final Report
Blake Benton
Ariel Thompson Cash
Vijay Karnatak
Jared Lee
Couper Turkewitz

Can the weatherman sell your house?

Introduction and background information:

The purpose of this project was to explore the relationship between weather and real estate transactions, more specifically this team analyzed the effect of mean average temperature and precipitation at the time a property comes onto the market has on the sales price and/or the time it takes to sell. This project falls under category 1: analytics, with homebuying being both a consumer practice and an investment strategy. In our research no other entities have performed this type of analysis. Previous research in the field of real estate has been focused on what buyers look for in a house based on the features of the property, but little research has been done to determine the influence the buying process can have on the purchase of a home. Recently, Hohenstatt and Kaesbauer conducted research based on google search engine queries for buyers to determine what people really want out of a home [1]. This study is pertinent to this group project because it emphasizes the importance of analytics on the real estate market. Analytics firms are focused on what features of a home would entice a buyer, but they fall short on looking into other factors that affect buying. For this project, we will focus on the buying environment, specifically the weather at time of purchase to determine if there are predictable changes in price or number of homes sold based on the weather. As the climate changes, weather patterns in most locations are going to change. Our model could be useful in starting the conversation around weather and homebuying. Some research is being done to look at the effects of extreme weather on the broader housing market, but this research does not consider individual transactions but may be useful when considering broad changes to the housing market [2]. Two studies have been found to support our use of analytics models in our study. First is research from March 2020 which investigated the use of big data in real estate [3], which is based in Australia. This study found that using big data, they could better understand buyer sentiment and anticipate customer regrets based on the lack of full information from the seller. This study suggests that there can be a third party other than the seller and real estate agents to provide full information about a property prior to purchase. The final and most consequential paper we researched is about machine learning and predicting housing prices with regression techniques [4]. This paper is the basis for our regression methodology. The papers discussed here provide a groundwork upon which we will build our models. While the existing research has considered process of homes based on house features and the markets have been investigated with respect to climate change, we have found no evidence of research into home pricing and selling based on local weather. Our research may be used to sell homes to buyers more effectively.

Hypothesis:

This group investigated three hypotheses: The first hypothesis states that as temperature increases, Median home sale price, median price per square foot and percent of homes that go off the market in two weeks will increase. Second, if the temperature is too hot or too cold, the sales price will decrease and the time it takes for real estate to sell will increase, therefore there will be some correlation between the deviation from an ideal temperature and

sales price. Additionally, we hypothesized that the more precipitation there is in the time the home is on the market, the longer it will take to sell; we also checked for a correlation between more precipitation and lower sales price, but we did not anticipate a strong correlation there. If our hypotheses are supported, for the average homeowner this type of knowledge could be useful in determining when to list their property for sale, and how to set their asking price. For example, if a homeowner needs to sell quickly, and knows that the weather is going to be very rainy for the next two weeks, they may preemptively set a low asking price to drive the smaller activity during the rain to their listing. Alternatively, if a homeowner is unconcerned with the amount of time it takes to sell, but is instead trying to maximize the sales price, they may want to wait for the weather to reach a certain temperature and have no rain forecast for the two weeks post listing.

Overview of the Data:

We will use two large datasets to develop a model. The first is real estate data from Redfin, a home-selling website, for major cities in the United Stated and includes information on median home price, price per square foot, if the home was sold within two weeks and the asking price of the home. The real estate data is aggregate data and is reported for each month. Since the data set was so large, we chose to take a subset of the data from four cities: Sacramento, Dallas, Atlanta and New York, which we determined included a broad range of climates and housing markets. We also included the type of home in our dataset. The second dataset is weather data from the National Oceanic and Atmospheric Administration, NOAA, and includes daily temperature and precipitation for major cities. There are several decisions that were made in combining these two datasets for use in analysis. Since the real estate data is from a company that makes money from their data, we were not able to see the data at the individual property level, rather the data is aggregated into monthly averages. Because of this date structure, we modified the weather data to be aggregate monthly values as well, which transformed the weather data into average monthly precipitation and average monthly temperature. Data was merged on date.

With the final merged dataset, cleaning was performed to remove N/A values from precipitation rather than impute values. In exploratory data analysis, we looked for correlation between factors to determine which predictors were likely to be significant in linear regression. Figure 1.1 shows the correlation matrix between values, which lead this group to conclude average monthly temperature (TAVG), average precipitation (PRCP), median sales price, median price per square foot, and percent off market in two weeks as notable and usable predictors for linear regression.

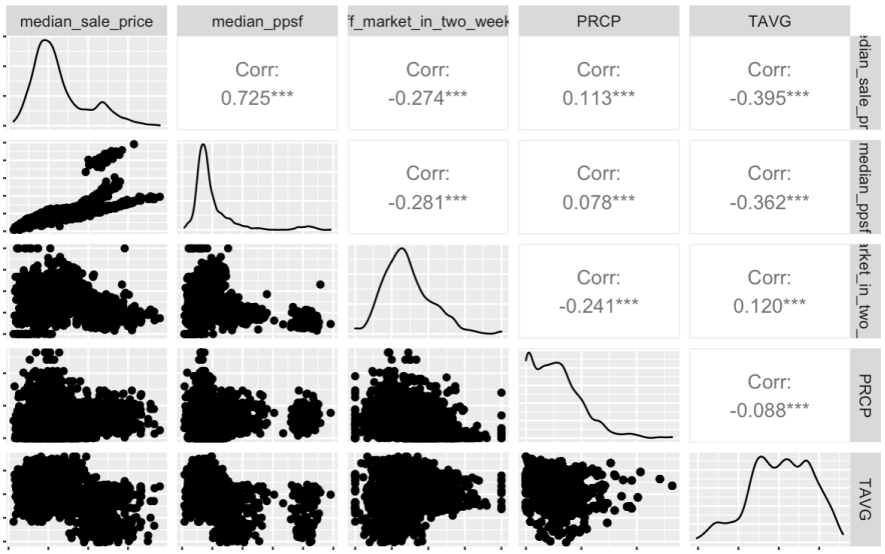

Figure 1.1 Correlation matrix. *** indicates significance to .001 p-value

Modeling Methodology:

Analysis by linear regression was chosen to determine the relationships between key variables, and we began the investigation by looking at the effect weather has on median sale price. As seen in figure 1.1, precipitation has a slight direct relationship with median sale price with a correlation coefficient of 0.113 and average temperature has a medium indirect relationship with median sale price with a correlation coefficient of -0.362. We started by building four models that looked at the relationship between median sale price and precipitation with median sale price as the dependent variable. To determine if the scale of our data was consistent, we experimented with log transformations on median home price and precipitation. The four models consisted of a linear-linear, linear-log, log-linear and log-log model. Since some of the precipitation values were zero, the log of precipitation + 1 was used. The adjusted R squared values are show in figure 1.2.

Model	**Adjusted R squared**
Linear-Linear	0.01225
Linear-Log	0.01496
Log-Linear	0.0119
Log-Log	0.01428

Figure 1.2 Adjusted R squared vales of log transformed variables in linear regression

The linear-log model had the best fit, however with a low adjusted R squared value of just 0.01496. This model implies that median sale price increases by $3,243.48 for each 1% increase in precipitation, holding all other factors constant. Because the linear-log model accounts for the most variation in the data, we chose it as our base model upon which we would add other factors. The model is shown below.

Median sale price = 350,578 + 324,348 * log(precipitation +1)

Next, we considered adding the average monthly temperature to see if it improved the model's fit. We built two models, one adding the average monthly temperature and one adding the log of the average monthly temperature. The adjusted R squared values are shown in figure 1.3. Adding the average monthly temperature improves the model with the Log of average monthly temperature improving the model the most with an adjusted R squared of 0.1931.

Model	Adjusted R squared
Linear-Log-Linear	0.1625
Linear-Log-Log	0.1931

Figure 1.3 Adjusted R squared values of linear regression of log transformed median sales price, precipitation and average monthly temperature

The model of median home price with respect to average monthly temperature and precipitation implies that median sale price increases by $2,191.26 for each 1% increase in precipitation, holding all other factors constant and that median sale price decreases by $30,413 for each 1% increase in the monthly average temperature. Accounting for the log adjustment in precipitation and average monthly temperature, the updated regression model is shown below.

Median sale price = 1,600,300 + 219,126 * log(precipitation +1) – 304,130 * log(monthly average temperature)

While this model is an improvement on the previous model, we still believed other factors influenced median sale price such as property type and city. The next factor we chose to incorporate into the model was the property type factor. Because a larger home would invariable cost more than a smaller home, we thought property type would take into account the differences based on property size and land value that would come with those properties. Property type is broken up into 5 different factors: All Residential, Condo/Co-op, Multi-Family (2-4 Unit), Single Family Residential and Townhouse. We added property type to the model using All Residential as the base case. This new model has an improved adjusted R squared of 0.2275. The model is shown below.

Median sale price = 1,604,149 + 218,778 * log(precipitation +1) – 304,139 * log(monthly average temperature) – 75,443 * Condo/Co-op + 41,984 * Multi-Family + 4,894 * Single Family Residential + 9,994 * Townhouse

When it comes to weather, this model implies that, on average, median sale price increases by $2,187.78 for each 1% increase in precipitation and decreases by $30,413.90 for each 1% increase in average monthly temperature holding all other factors constant. In regard to property type, this model implies that, on average, holding all other factors constant, Condos/co-ops are $75,443 less than All Residential properties, Multi-Family units are $41,984 more than All Residential units, Single Family Residential units are $4,894 more than All Residential units and Townhouses are $9,994 more than All Residential units.

With full weather and property type accounted for, the group then added city to the model. The four cities in our data are Atlanta, Dallas, New York and Sacramento. Using Atlanta and All Residential as the base case we ended up with a model with a much-improved adjusted R squared of 0.7403. The model is:

Median sale price = 167,482 + 12,048 * log(precipitation +1) + 30,886 * log(monthly average temperature) – 75,443 * Condo/Co-op + 41,984 * Multi-Family + 4,894 * Single Family Residential + 7,870 * Townhouse + 10,764 * Dallas + 439,072 * New York – 2,977 * Sacramento

Including the city in the model reduces the effect that weather has on median sale price. Now median sale price, on average, only increase by $120.48 for each 1% increase in precipitation holding all other factors constant. Also, temperature now has a direct relationship with median sale price. Median sale price, on average, increases $308.86 for each 1% increase in average monthly temperature holding all other factors constant. Property type stays mostly the same with the only change being that a townhouse on average, is $7,870 more than All Residential property types, holding all other factors constant. When it comes to city, we notice that, on average, holding all other factors constant, median sale prices in Dallas are $10,764 more than Atlanta, $439,072 more in New York than Atlanta and $2,977 less in Sacramento than in Atlanta.

With the model having incorporated all of the factors chosen from the correlation matrix in figure 1.1, significance of the factors needed to be determined. Using a significance threshold of 0.10, we see that the null hypothesis that the coefficient is zero cannot be ruled out for the factors log(precipitation +1), Single Family Residential, Townhouse and Sacramento. Figure 1.4 are p values for the coefficients of the model.

Coefficient	**P value**
Intercept	4.54e-05
Log(precipitation +1)	0.71666
Log(monthly average temperature)	0.00137
Condo/Co-op	< 2e-16
Multi-Family	5.17e-09
Single Family Residential	0.49423
Townhouse	0.27451
Dallas	0.09070
New York	< 2e-16
Sacramento	0.67517

Figure 1.4 p-values of all factors regressed on median home price

Based on p-value significance, we removed precipitation as a variable and reduced the property type variable to three factors: Condo/Co-op, Multi-Family and all others. We also reduced the cities to three factors: New York, Dallas and all others. The following model with an adjusted R squared of 0.7406 and a base case of all other cities and all other property types, is the final model for this metric.

Median sale price = 170,885 + 30,991 * log(monthly average temperature) – 79,683 * Condo/Co-op + 37,744 * Multi-Family + 12,421 * Dallas + 440,994 * New York

In this final model for median sale price, we see a few things, on average, given all other factors constant. A 1% increase in average monthly temperature increases median sale price by $309.91. Condos/Co-ops have a median sale price $79,783 less than all other properties besides Multi-Family and Multi-Family units have a median sale price $37,744 higher than all other properties besides Condos/Co-ops. Median sale prices in Dallas are $12,421 higher than in Sacramento and Atlanta and median sale prices in New York are $440,994 higher than in Sacramento and Atlanta.

Using the same methods explained above, the team also performed linear regression analysis on median price per square foot and the Percent of home off market within two weeks. The rationale for these two metrics is that the price per square foot would further help to standardize the size of the property, and the percent of homes off the marker in two weeks would give us an idea the difference weather can make in the speed with which a property is sold. The model for median price per square foot had an adjusted R squared value of 0.632. The final model for median price per square foot is below.

Median price per square foot = 168.86 + .37 * monthly average temperature + 100.93 * Condo/Co-op -63.21 * Multi-Family – 37.34 * single family - 56.04 * Townhouse -13.22 * Dallas + 353 * New York + 13.32 * Sacramento

Using a significance threshold of 0.10 we were unable to reject the null hypothesis that the coefficients of precipitation and monthly average temperature are zero for the percent of houses off the market within two weeks model. Therefore, we determined that there is no relationship between quickness of sale and weather when considering property type and city factors.

Discussion and key visualizations:

The implications of these models are best represented graphically. Figure 2.1.1 displays the inverse relationship between average monthly temperature and median sales price. The negative slope of the graph suggests that an increase in temperature will decrease the value of the property. Therefore, average monthly temperature alone leads to an incomplete model. With the addition of home type in figure 2.1.2, the inverse relationship persists. Even accounting for home type, the graphs would suggest that the hypothesis is unsupported, and that home price decreases with increased temperature. This team considered these outcomes and considered whether there is an influence of the broad range of temperatures seen in New York versus Dallas that may be accounting for the negative trend.

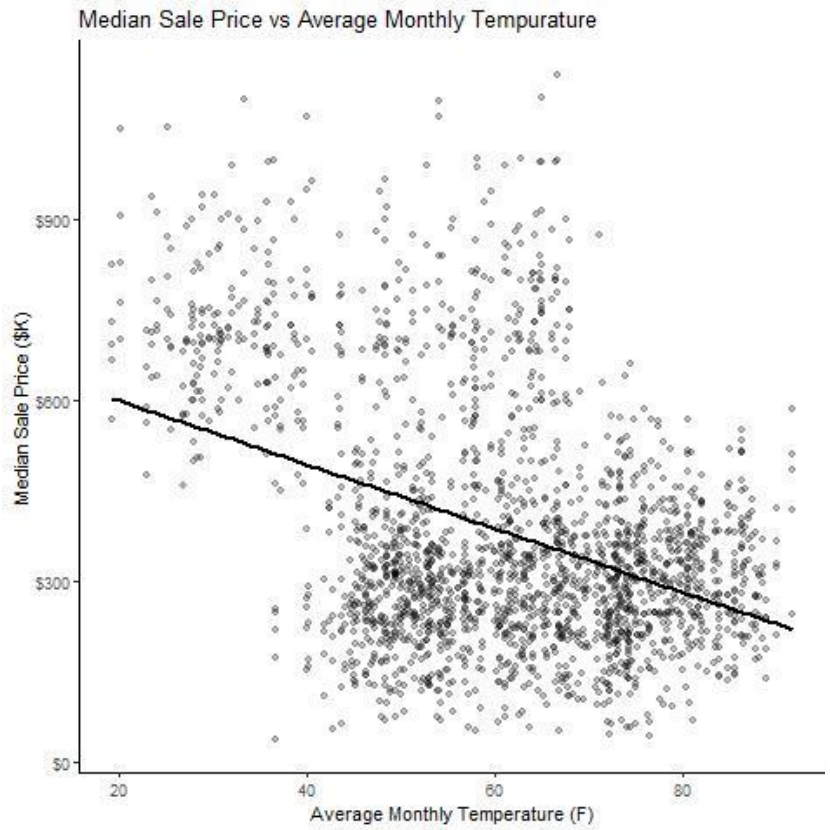

Figure 2.1.1 Average monthly temperature versus sale price alone

Figure 2.1.2 Average monthly temperature versus sale price and property type

When the difference in city is accounted for by adding them to the regression model, each of the four selected cities shows its own positive relationship between median home price and average monthly temperature, as seen in figure 2.2. This figure also highlights that New York is, on average a colder city, and Dallas tends to be much warmer, but the positive trend holds true. The same positive trend is evident when considering the size of the home. Figure 2.3 shows the relationship between average monthly temperature and price per square foot.

Figure 2.2 Average monthly temperature versus sale price based on property type and city

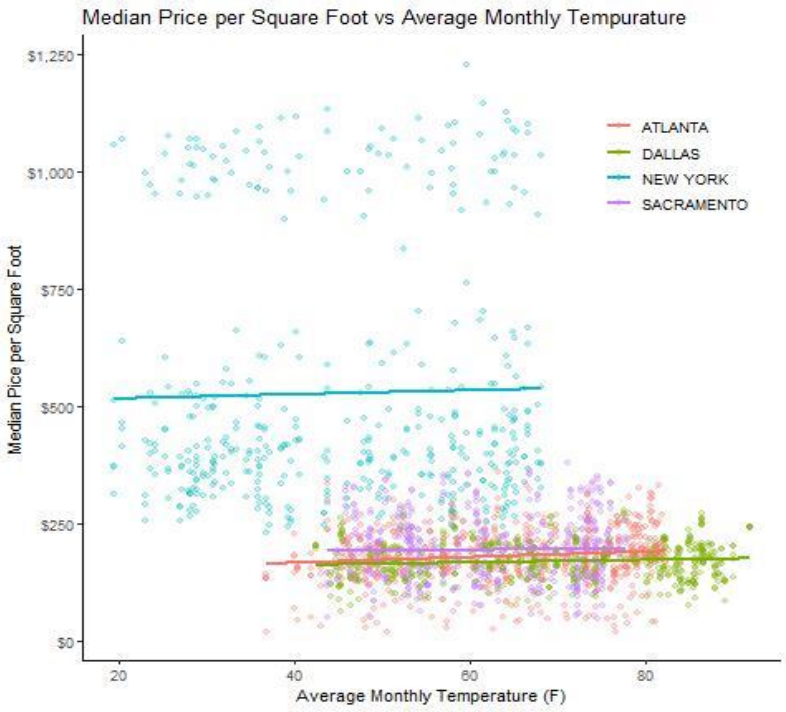

Figure 2.3 Average monthly temperature versus price per square foot alone and based on city

The final key figure that helps to shape a clear picture of the effect of warmer weather on the median price of a home can be found in figure 2.4. While it is common understanding to real estate agents that more homes sell in the spring and summer, we found that it is not true that homes sell for more in these warmer months. By modeling the median home price per month, we have found no cyclical or seasonal pattern to home price, and have therefore determined that price is truly influenced by temperature, and not simply a cyclical pattern of seasons.

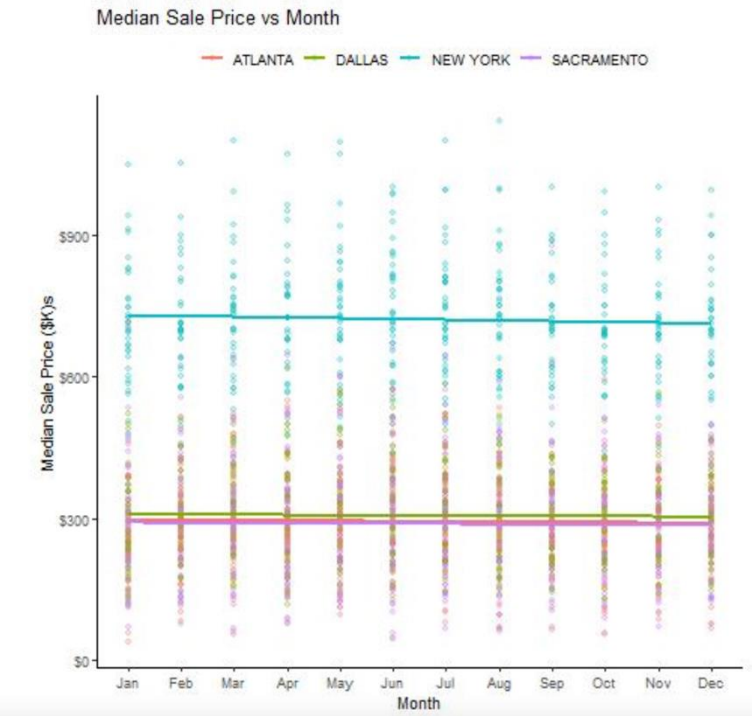

Figure 2.4 Median sales price by month in 4 cities

Challenges:

There were some challenges associated with the completion of this analysis. Regarding the data, weather data is not normalized for location, so there may be trends that resulted from cities having broader or narrower ranges for temperature. In future studies the team agrees that a normalization of the temperature for each city would be a strategic way to account for these differences. Additionally, rather than considering raw temperature, we wondered if there is an optimal temperature at which homes sell for the greatest median sale price, and could we measure deviation from the highest sales price temperature, which would be more predictive. Another challenge was with the real estate data. Because Redfin is itself a business, the publicly available data is aggregated monthly while the data is given daily on NOAA. Listing-level data would require MLS scraping which time constraints do not allow for this project, but would be interesting for further research.

Conclusions:

The results of this analysis would suggest that the factors that cause the largest difference in home price are location and property type, but these factors are features that homeowners cannot change about their house. There is, however, a small but significant difference that average monthly temperature makes in the price of the home. The final model

suggests a 1% increase in average temperature increases the price of a home by $310. For example, waiting for the temperature to rise from 70 to 80 degrees would increase the sales price by nearly $4,430. The implications for this study are that people who are selling a property would be best advised to wait until warmer weather to sell a home and buyers should wait until colder temperatures to purchase a home. While the differences in price are small to a single transaction, these models can be extrapolated for large investment firms and used to further profits of those companies. Indeed, if Redfin were to include weather data in their business practices, the company would be far more profitable. The implications of this study could be large to a company who moved real estate property on a large scale.

Reference:

1. Ralf Hohenstatt & Manuel Kaesbauer (2014) GECO's Weather Forecast for the U.K. Housing Market: To What Extent Can We Rely on Google Econometrics?, Journal of Real Estate Research, 36:2, 253-282, DOI: 10.1080/10835547.2014.12091387
2. Hirsch, J., Braun, T. and Bienert, S. (2015), "Assessment of climatic risks for real estate", Property Management, Vol. 33 No. 5, pp. 494-518. https://doi.org/10.1108/PM-01-2015-0005
3. Munawar HS, Qayyum S, Ullah F, Sepasgozar S. Big Data and Its Applications in Smart Real Estate and the Disaster Management Life Cycle: A Systematic Analysis. Big Data and Cognitive Computing. 2020; 4(2):4. https://doi.org/10.3390/bdcc4020004
4. Dange, T., Mishra, A., Jagtap, A., Chavhan, S., & Chavan, N. (2022). Machine Learning based House Price Prediction using Regression Techniques.