HOUSE PRICE PREDICTION

December 7, 2020

I. TOPIC

Changing jobs and moving cross-country are two of the most stressful and impactful decisions a person can make in their lifetime. When both of these life changes are undertaken at the same time, it is easy for anyone to become overwhelmed with decision paralysis in the face of so many factors to consider: salary, commute, cost of living, distance from family, location, property value, and more. Today, we have more data than ever before at our fingertips that can help us to make sense of these factors and come to a decision based on solid data analysis. Data analysis can help remove some of the emotion that often clouds the consideration of major life changes. When presented with a clear and concise collection of unbiased data, one can consider what the data supports and then decide on a course of action with a good understanding of all the information. This is not to say that emotion will never impact decisions, but having a solid understanding of the data may help lead to a more level-headed choice.

We consider the case of a client from our professional consultancy practice: a junior software developer who wants to know whether it is a good idea to move from Seattle to a suburb of Atlanta, GA based on a prestigious job offer from Microsoft. He is chiefly concerned with two factors: real estate value and commute. He currently owns a home in the 98007 zip code of Seattle, which is worth (x). His current salary as a software developer at Expedia is (x). The offer from Microsoft is for a senior software developer position with a salary of (x). The Georgia Microsoft office is located in Alpharetta and our client wishes to buy a home no more than ten miles from the office if he chooses to accept the offer. We have selected (x) zip codes within this radius for his consideration. Using data from the renowned real estate website Zillow, we are able to determine current real estate price trends in both Seattle and Alpharetta, as well as predict future values. These factors, combined with the commuting distance, will help us make a recommendation to this client.

II. DATA COLLECTION

Zillow launched in 2006, coincidentally the current home to our client. It has since grown to be the leading real estate database in the United States, containing data on over 110 million homes for sale, for rent, and even those that are not on the market at all. It even contains data dating all the way back to

1996, ten years before the company was even formed. The Zillow Home Value Index, or ZHVI, is the company's "flagship measure of both the typical home value as well as housing market appreciation currently and over time." (Zillow) Data is published monthly and considers both new construction homes and homes that have not been on the market for years. This provides a more comprehensive view of overall market trends because Zillow's data does not only rely on home sales records from any given period. Zillow's data also provides for a superior level of granularity, giving users the option to observe trends for very small regions or specific subsets of homes. (Zillow)

The base of the ZVHI is Zillow's signature Zestimate, which denotes the company's estimated value for any given home. The Zestimate uses both public data and user-submitted data while also accounting for factors such as market condition and location. Zillow uses this data to formulate millions of statistical and machine learning methods that can examine data points for each home. (Zillow) A Zestimate's accuracy is determined by comparing the final sales price to the Zestimate that was active on the date of sale. The Zestimate is within 10% of the final sales price more than 95% of the time, with a nationwide median error of 1.9%.

In 2019, Zillow made some changes to its proprietary ZHVI algorithm to improve its accuracy. In its previous form, the ZHVI calculated median Zestimate value for a fixed set of homes over a given period. This value was meant to represent the median home value of the given area. Appreciation of these values could be taken to be total market appreciation, without consideration to more expensive homes driving up the median by selling at higher prices. With the 2019 improvement, a weighted mean of appreciation is used, which weights each home in the index proportional to the value of its Zestimate in previous months. This provides a better calculation in which appreciation of the index can be interpreted as appreciation of the total market. The current model also corrects for home appreciation driven by home improvement projects, because such improvements are not indicative of total market appreciation. Zillow's model represents the value of such homes as if the improvements had not taken place.

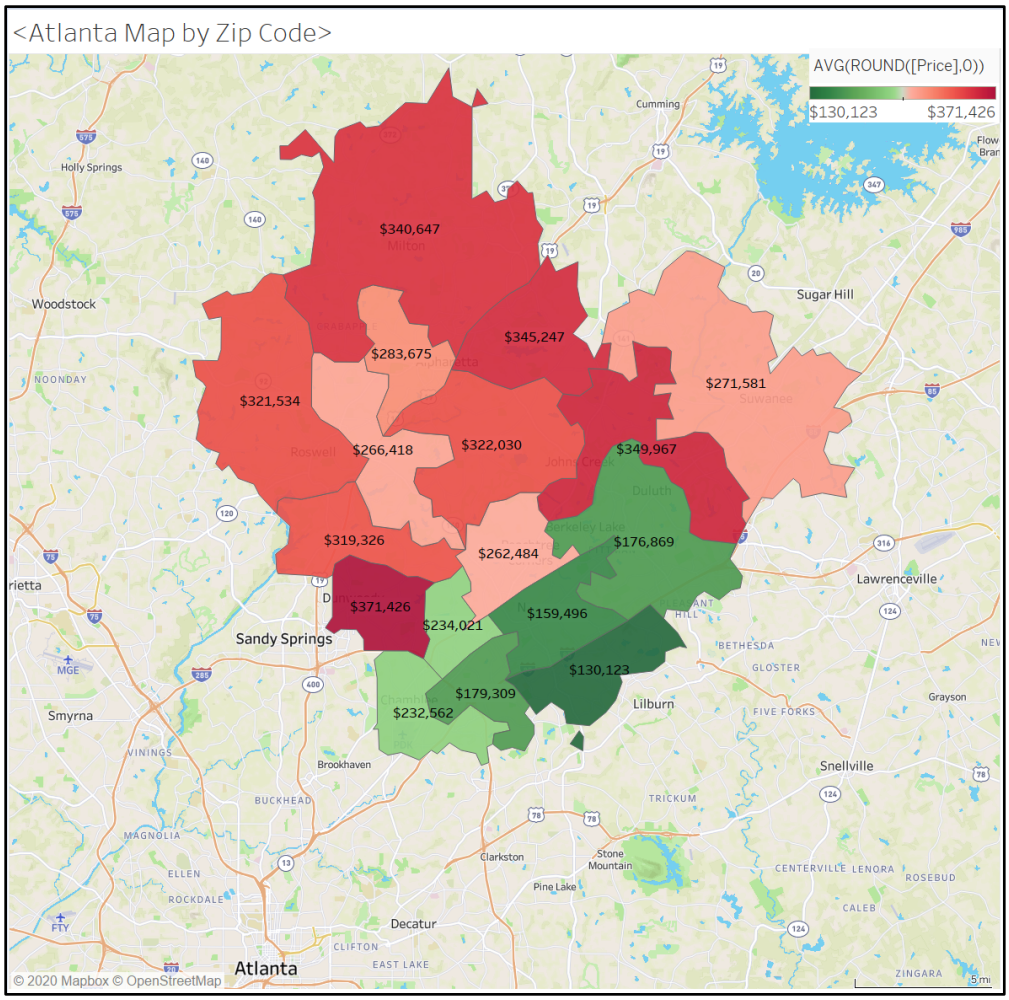
Zillow's research arm is separate from the revenue-generating center of the company. It is dedicated to creating timely, accurate, unbiased data models using a combination of public and user-generated data along with its proprietary statistical and machine learning models. The Zillow Research branch aims to be transparent about data issues and will never alter data to manipulate results. This commitment to empowering consumers and the highly accurate results of Zillow's models mean it is a deep and rich data source that can be trusted.

III. ANALYSIS

A. Descriptive

As part of descriptive analytics, we intended to analyze the various zip codes in Alpharetta, GA and observe the trends. Alpharetta, a suburb of Atlanta, is the location of the Microsoft offices. The below visualization displays the graphical representation of the average home prices of the zip codes, which are in about 10 miles radius from the work location of our client. It has highlighted the zip code as dark green for the most affordable home prices, and dark red for the zip code with the most expensive home prices.

Figure 1



When the data was displayed graphically for each year, an overall positive trend in home prices can be seen between the years 1996 - 2020. The trends of all the zip codes reflect a similar pattern, wherein a drop in average home prices can be seen since 2008 - 2011. This reflects the Great Recession, which began in 2008 and culminated in a massive housing market crash. From 2012, the housing market

saw a gradual recovery, where a gradual rise in average home prices can be seen from year 2012 onwards.

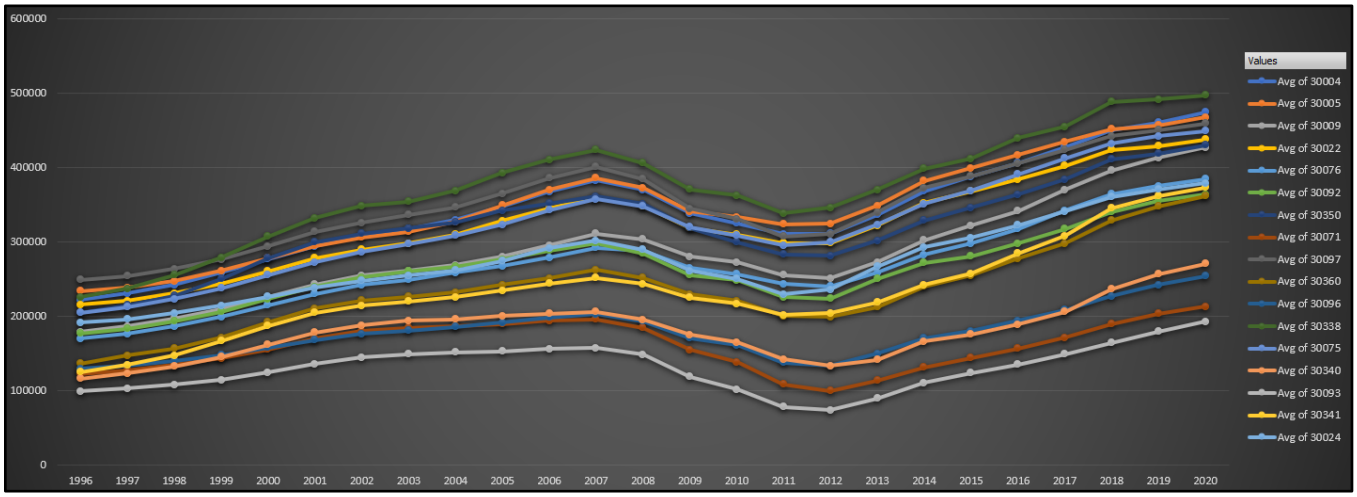


Figure 2

For the purpose of our case study, our client is evaluating zip codes which are within a 10-mile radius from his new job location. The following zip codes were selected by our client, considering the distance from work and average selling price of homes.

Zipcodes	30009	30022	30076	30092	30005	30350	30071	30097	30360	30096	30338	30004	30075	30340	30093	30341	30024
Miles	0	0.62	3.1	4.3	4.9	4.9	6.2	6.2	6.2	6.2	6.8	7.4	7.4	8.6	9.3	9.3	9.3
Average Home Price	283675.2	322030	266417.8	262484	345246.9	319326.2	159495.7	349967	234021.4	176869.3	371425.8	340646.9	321534	179309.5	130123.4	232562.3	271581.3
			Closer yet affordable				Average driving distance,yet affordable	Average driving distance,and expensive			Most expensive				Farthest and most affordable		

Figure 3 below shows the average selling prices of the zip codes selected, where the line indicates the most expensive and most affordable zip codes, and columns indicate the trends and range of the most expensive and most affordable zip codes that our client is evaluating.

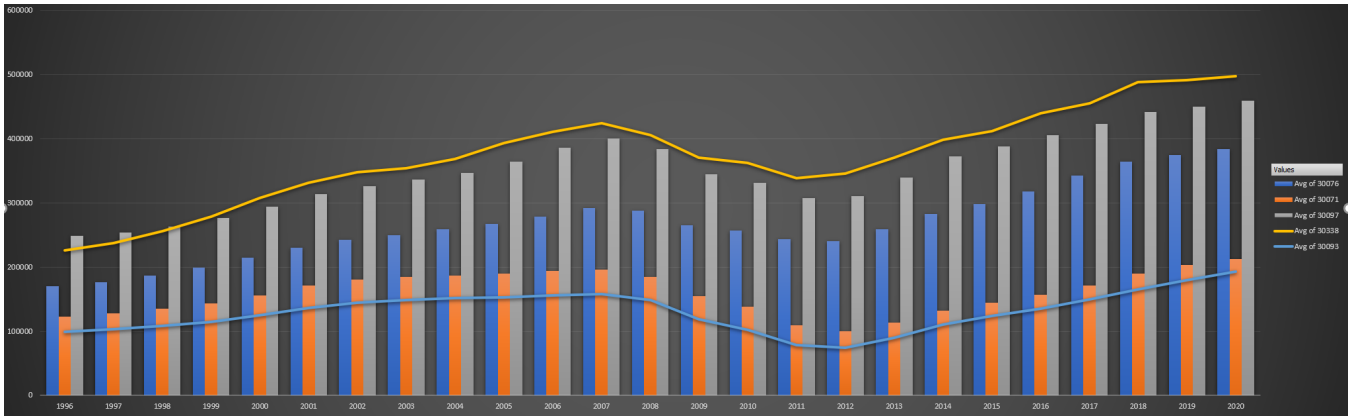


Figure 3

B. Predictive

The Great Recession, which began in 2008, was precipitated in large part by the subprime mortgage crisis, in which banks were providing incredibly risky home loans to consumers who could not afford to repay them. This culminated in a massive housing market crash as millions of homeowners foreclosed upon homes for which they could not pay. After this crisis, the United States passed bank oversight legislation to prevent a similar situation from happening in the future. This scenario was an aberration in the US housing market, not a regularly occurring event. Hence, our predictive model takes median home sales prices for each zip code starting in 2012, after the US began to recover from the Great Recession. We found that the recession home values were causing the predictive model to take on an inaccurate cyclical pattern. For this reason, we decided to remove years before 2012 from the training set in order to better train the model. This has given us 7 years, or 84 months, worth of data with which to predict a trend.

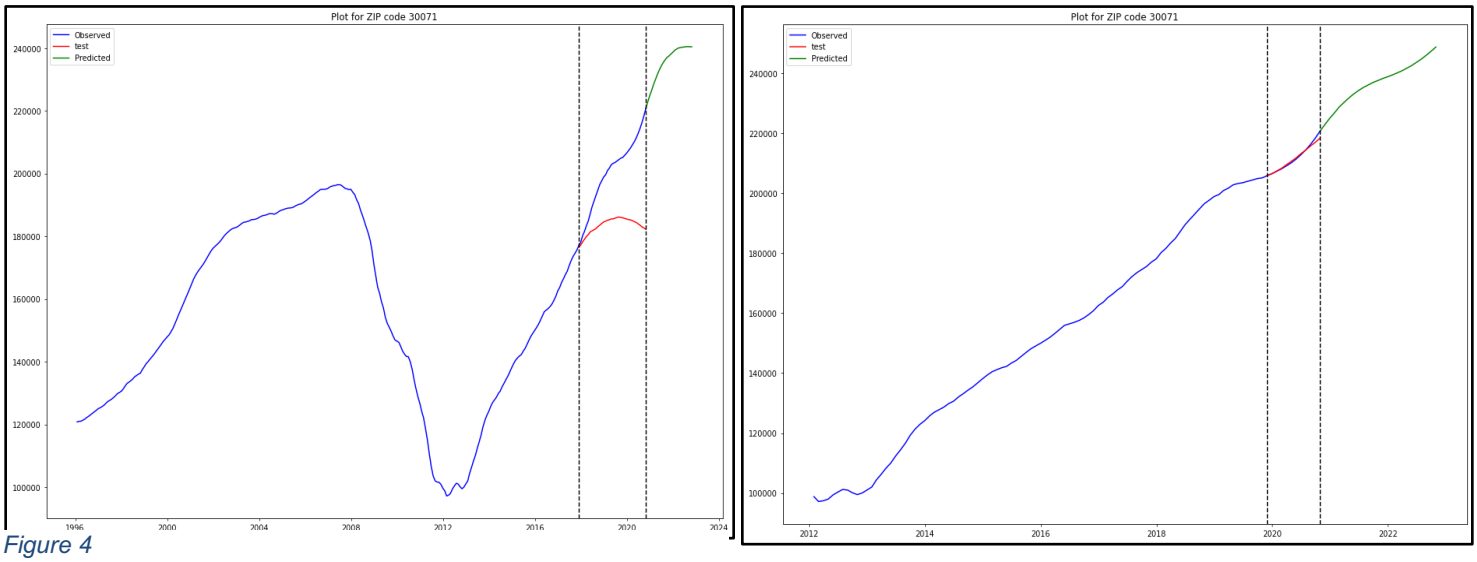


Figure 4

Our initial prediction model was producing MAPE values as high as 18.76%, meaning our forecasted home value predictions could have been off by almost 19% from actual values. The visualizations in Figure 4 below represents the initial predictive model versus the predictive model which excluded home values before 2012 from the training set. The model on the left predicts test values that would be indicative of a cyclical pattern, because the training set has taught the model to expect another devastating housing crash as seen during the Great Recession. Predicted values would show median home price decreasing over the next year, which is inconsistent with the trend seen in the data. The model on the right eliminates that data from the training set, therefore producing a more accurate predictive model with a MAPE value of only 4.6%.

While the post-2012 models for each zip code were better overall than the models using data beginning in 1996, each zip code has a different MAPE value, which is an indication of how different areas of the country experienced the housing market recovery differently. Places where the market recovery was more volatile will be harder for the linear model to predict, even with the elimination of Great Recession-era values, because the training data shows a more inconsistent trend. The zip code 30338, seen in Figure 5 below, clearly had a more turbulent recovery period than the zip code 30071, seen in Figure 4. The zip code 30071 shows a steady upward trend with no significant drops. The zip code 30338, however, displays an almost seasonal trend combined with a positive linear trend, where the market climbs, then experiences small drops, and then recovers and continues to climb.

Once again, the model on the left shows a tendency toward a cyclical pattern because the training

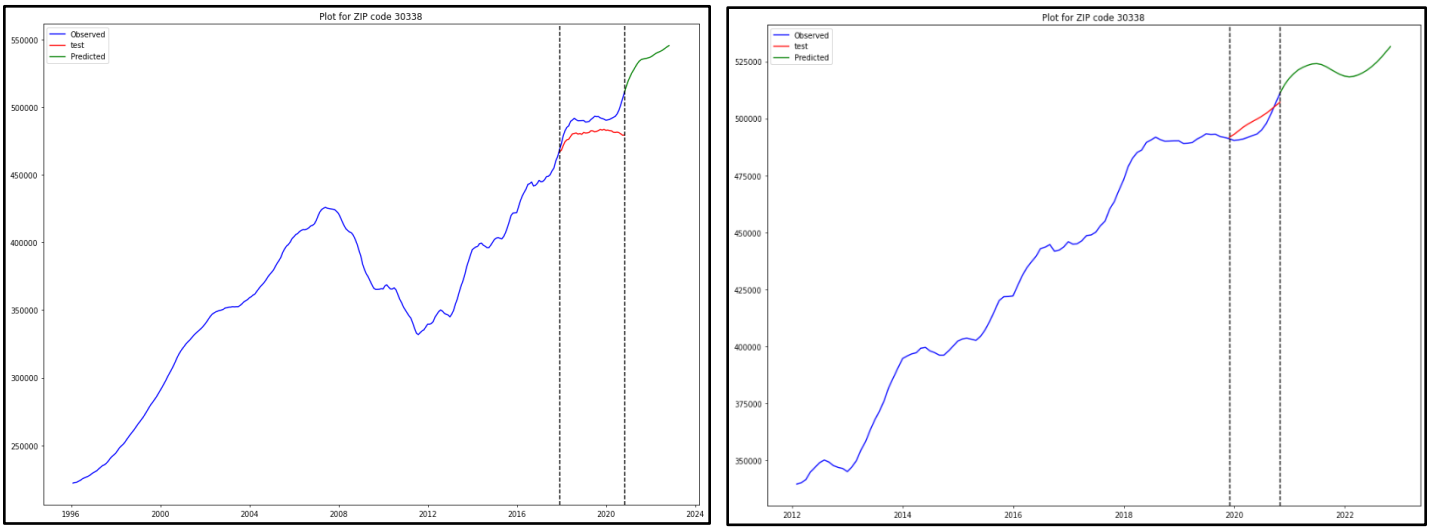


Figure 5

data has taught the model to expect another market crash. This resulted in a MAPE value of 11.45%, compared to the post-Recession model with a MAPE value of merely 1.64%.

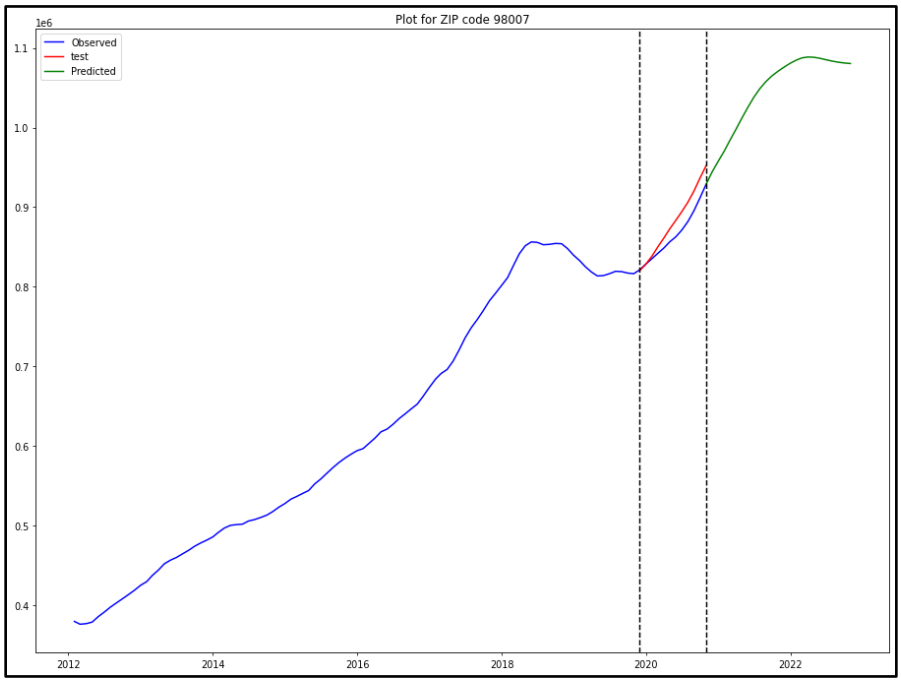


Figure 6

We also consider the predictive model for our client's current home in Seattle. The housing market for zip code 98007 experienced a drop around mid-2019 before recovering through 2020. This has caused a slight error in the predictive model compared to observed values, but with a MAPE value of 7.75%, the model is still a good predictive indicator of home value over the next year. The model may be more erroneous for predictions after 2022, because it anticipates another market decrease as seen in 2019. For the purposes of advising our client, who would like information about home values in the next 12 months, the model is a good fit.

C. Prescriptive

As discussed, the decision to move across the country and change jobs can be a very stressful time. It is hard to decide based strictly on outside data because these choices do have an emotional component as well. For this reason, we have chosen to employ a utility theory optimization algorithm. Utility theory considers the decision-maker's tolerance for risk. In this case, the decision-maker is our client. By taking his tolerance for different options into account, we can quantify some of the key emotional factors of making this choice and account for the human variable in the decision. Every person is different; some are more motivated by money, some by quality of life, some by job fulfillment.

We first utilized a binary linear optimization model, as shown in Fig. 7, to determine the best Georgia neighborhood to live in based on cost, taking into account current home values, projected

increase, property taxes, commuting costs, and standard utilities. All values for the home value increase, property tax, and utility costs are reflective of typical rates in Georgia for the respective zip codes. The decision variables for this model are each of the five zip codes, while the constraints are that we must have non-negative values for the variables, and that our client wishes to see the home value increase at least 1 percent over the course of a year. The objective function is to minimize expenses. This model has determined that the optimal zip code in the chosen selection is 30071, on a strictly monetary basis.

Atlanta Taxes (filing for single)			
	Marginal Tax Rate	Effective Tax Rate	Taxes
Fed	24.00%	17.75%	$ 24,850.00
FICA	1.45%	7.34%	$ 10,276.00
State	5.75%	5.33%	$ 7,462.00
		30.42%	$ 42,588.00

Salary
$ 145,000.00

Take Home Money
$ 94,461.24

Zip Code	2020 Median home price	1 Year Increase	Price value	Property tax %	Annual Prop. Tax
30076	$ 384,256.60	4.0400%	$ 399,780.57	1.0270%	$ 3,946.32
30071	$ 212,877.40	6.8900%	$ 227,544.65	1.2190%	$ 2,594.98
30097	$ 458,998.30	0.8000%	$ 462,670.29	1.0270%	$ 4,713.91
30338	$ 497,108.30	0.8700%	$ 501,433.14	1.0940%	$ 5,438.36
30093	$ 192,974.90	8.4600%	$ 209,300.58	1.2190%	$ 2,352.36
		6.8900% => 1.0000%			

Dec Variables	
Z1	0
Z2	1
Z3	0
Z4	0
Z5	0
	1 = 1

Zip Code	Miles	Commute time (min)	Drive Work Days	Total Miles	Annual Fuel Money
30076	3.11	15	260	1615.57	$ 807.78
30071	6.21	26	260	3231.13	$ 1,615.57
30097	6.21	15	260	3231.13	$ 1,615.57
30338	6.84	20	260	3554.24	$ 1,777.12
30093	9.32	31	260	4846.70	$ 2,423.35

Minimize Costs	
Obj Func	$7,950.76

Zip Code	Natural Gas	Electricity	Water	Internet	Annual Prop. Fees
30076	$84.43	$107.42	$59.83	$60.00	$3,740.22
30071	$84.43	$107.42	$59.83	$60.00	$3,740.22
30097	$84.43	$107.42	$59.83	$60.00	$3,740.22
30338	$84.43	$107.42	$59.83	$60.00	$3,740.22
30093	$84.43	$107.42	$59.83	$60.00	$3,740.22

However, there are many other factors to consider when undertaking such a life-changing decision.

We have worked with our client to determine the following main factors that will affect his decision: salary, commute time, projected home value increase, violent crime, property crime, moving

Binary Linear Minimization Model:

Setup

Let Z be a binary variable representing 5 selected zip codes

- Z1 = 1 --> zip code chosen to live in
- Z1 = 0 --> otherwise

Let P be price of property tax for 5 selected zip codes

Let F be price of fuel costs for 5 selected zip codes

Let U be price of utility costs for 5 selected zip codes

Minimize

(Z1*P1 + Z2*P2 + Z3*P3 + Z4*P4 + Z5*P5) +
(Z1*F1 + Z2*F2 + Z3*F3 + Z4*F4 + Z5*F5) +
(Z1*U1 + Z2*U2 + Z3*U3 + Z4*U4 + Z5*U5)

Subject To

(Z1 + Z2 + Z3 + Z4 + Z5) = 1

1 year property increase > 1.00%

All variables are nonnegative

costs, and amount of school spending per child (the client has three school-age children). Since our client does not want to work more than 10 miles from the prospective new job at Microsoft, we have only considered neighborhoods within that radius. Our client considers each additional hour of commute to be worth half an hour's wage, when salary is broken down to an hourly amount, assuming a 40-hour workweek. He also considers both violent and property crime to be high risk factors. We have developed the following utility function to account for his needs:

$$Utility = Salary - \frac{Salary}{2 * 40 * 52} * (2 * 5 * 52 * \frac{commute\ tine(minute)}{60}) + \frac{Salary}{3} * \%HouseValueIncrease + \$School\ spending\ per\ student * 3\ kids - 1000 * (2 * violent\ crime + property\ crime) - Moving\ Cost$$

Zip code	Salary	1 Year Home Value Increase	Money schools spent per student	Commute Time(minutes)	Violent Crime	Property Crime	Moving Constant	Utility
98007	140000	0.0840	14197	5	19.4	31.1	0	115153
30076	145000	0.0404	10734	15	28	35.5	8000	75123
30071	145000	0.0689	9550	26	43.2	52.8	8000	21926
30097	145000	0.0080	10734	15	12.6	18.1	8000	121757
30338	145000	0.0087	9451	20	18.2	31.6	8000	91732
30093	145000	0.0846	9550	31	56.8	66.9	8000	-20126

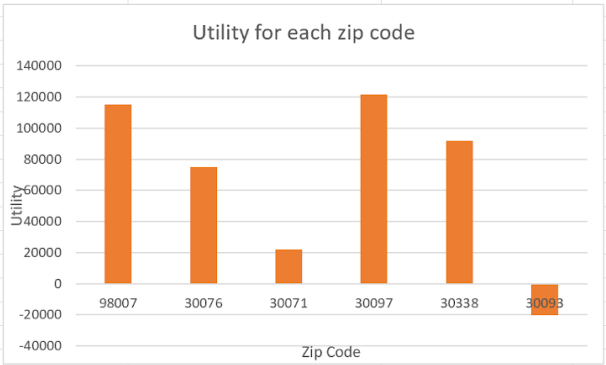


Figure 8

```
Max utility value: 121757.41666666666

98007: 0.0

30076: 0.0

30071: 0.0

30097: 1.0

30038: 0.0

30093: 0.0
```

Figure 9

If our client wishes to consider only salary and home value, the choice is simple: he should stay in Seattle at his current job. When we employ a utility algorithm considering all the client's desired decision factors, the conclusion becomes less obvious. The results of the utility algorithm can be seen in Fig. 8 above. When we place this utility function into a Python optimization function as seen in Fig.9, we see a new choice of zip code: 30097. The median values for the client's current home zip code and a home in the 30097 Georgia zip code are very close, but the Georgia zip code utility is slightly higher, owing mainly to lower crime rates in the area compared to Seattle. The Seattle zip code beats the Georgia zip code in every other factor, yet the client's much lower tolerance for risk in crime rate shifts the utility value in Georgia's favor.

IV. CONCLUSION

This client's question is a great example of the need for multiple forms of analysis to form a solid viewpoint. If we were to consider only the linear optimization model, we might be recommending a suboptimal course of action for our client. The linear model alone does not consider the personal emphasis our client places on factors like crime rates, commute times, or school spending. Therefore, we have also considered the utility model to present a more personalized recommendation to the client with regard for his needs. Of course, a single model can never account for every factor. There are many other things which may influence the decision, such as the working environment, opportunities for advancement, or city lifestyle. There are other opportunity costs to consider such as missing time with extended family. We must also account for some of the limitations of our data source. Zillow compiles median home data, so exact home sales values for each home are not available. There are bound to be homes valued at much higher and much lower than the median in any given zip code. We can see an overall trend for an area, but we cannot predict with precision the exact value of any given home. Data is

a powerful tool for making a decision, but when it comes to things that affect one's personal life, it can only be one piece of the puzzle.

Works Cited

1. Zillow Research on Dec. 18, 2019. "Zillow Home Value Index Methodology, 2019 Revision: What's Changed & Why." *Zillow Research*, 20 Dec. 2019, www.zillow.com/research/zhvi-methodology-2019-highlights-26221/