UCONN
SCHOOL OF BUSINESS

OPIM 5770
CAPSTONE PROJECT REPORT

Analysis of Basement Systems Lead Data

Group 6
Aleksandra Kaminska
Abhinav Suggula
Jiancheng Shangguan
Qichen Hu
Samarpita Chakrabarty
Wei Wang
Yueh Chang

Contents

1. Executive Summary

In a capitalistic world, we know that there is a lot of saturation in many markets. The fundamental housing-needs market is no different. Basement systems is a leading provider of consumer needs for home repair in all aspects, especially on the east coast. There is a lot of potential for the company to expand its growth and provide its services to regions in which their services are not yet provided. There are also opportunities for new lines of business which will complement the potential growth of Basement Systems associations.

The purpose of the project is to provide Basement Systems some insights into their current state of business by analyzing the company data provided, identifying factors like climate and demographics which can potentially influence the demand for services offered by Basement Systems, and providing meaningful recommendations based on the analysis. Ultimately, the analysis is to help Treehouse Internet Group improve quality of the services that they provide and perhaps expand to new areas.

Both internal and external data was analyzed to find a potential correlation between various factors and the demand for services offered by the Basement System's association, which in turn, determined the increase in the number of leads. The trend and seasonality of the demand were also studied. Additionally, sentiment analysis has been performed on the customer reviews for different services associated with Treehouse. Some recommendations have been provided based on these findings.

2. Business Problem

The primary objective of the project is to identify the factors that qualify a good lead. This could mean many different things to a variety of companies. The lead data that was

provided by Basement Systems and Treehouse Marketing Group has already been filtered to consist of only good leads according to the sponsor's definition. This analysis is intended to help Treehouse Internet Group and its various associations acquire a large number of leads similar to those whose data was provided, by following the recommendations that are set out in this paper.

3. Background

Basement Systems started in 1987 and has rapidly grown since then. At first the company mostly only provided services for basement needs such as waterproofing, finishing, and so on. As time progressed, other associations came into the market mix and the Treehouse Internet Group was formed. Now Basement Systems being the flagship company only holds about 20% of that market mix to make way for all other affiliated associations. Not only does Basement Systems focus on basement needs, but with the new lines of work that have been added over the years, most structural needs to homes can be satisfied with Basement Systems.

4. Methodology

4.1. Obtaining Data

The initial challenge was understanding what outside data could be relevant information when it came to home needs. No one in our team is a homeowner so the search was primarily based off logical assumptions and conversations with the professor and sponsor.

During the initial search for external data the following information was looked for:

- Weather
    - Temperature
    - Flat lands

    - Soil
    - Precipitation
    - Wind
- House age
- Bugs, mice, rodent populations
- House values with and without finished basements
- Urbanization of areas
- Competitor presence
- Supply and demand of home building and repair needs

Unfortunately, not all the above information was available and certainly none that was available in Canada, therefore, the analysis has been confined to the United States.

Weather data was the easiest to find given its abundance and regular tracking throughout the entire USA. Home data was a bit more challenging. Even when information arose, it rarely gave specifics of the house which pertain to the services of Basement Systems. For example, housing data with the values of those houses could be found, rarely did it state, when the roof was last repaired or replaced, and if there was a crawl space, or how old the fence is which surrounds the perimeter, or if there even is a fence on the property.

The next challenge was finding data that pertained to the same time frame in which the analysis was being conducted. Weather is tracked regularly but many other components are not. Some studies and information gathering may have happened in 2010 regarding home repair, yet soil data was found which would only pertain to the last 4 years. Thus, much of the external data which was believed to be valuable, had to be discarded to keep a relatively similar time frame between all the other variables.

4.2. Data Preprocessing

In order to obtain a clean and correct dataset which could be analyzed, modeled and visualized to support the initial hypothesis and translate into actionable recommendations, detailed preprocessing was involved. There were four major steps taken to clean and merge the tables which included “Updated 10 Years.csv”, “Zip Code.csv” provided by sponsors and several external tables found on open data sources. All preprocessing and merging work was done using the software Python.

The first step involved deleting the columns or rows in the two datasets provided by the sponsor based on the following factors:

- If the columns which had a missing value rate higher than 95%.
- If the columns had more than 95% of duplicate records were deleted.
- If the rows had duplicated zip code information.

Secondly, the datasets were joined into a single table using zip code as the primary key. It is to be noted that only United States data was filtered for evaluating the sponsor's performance in the US market. Due to a lot of missing data both in the lead and, as previously mentioned, external data from Canada, it has been kept out of scope for the purpose of this project.

Next, columns such as “Year”, “Month”, and “Region” were created for the purpose of modeling and visualization. This allows us to better view our data in a comprehensive manner from different perspectives.

Lastly, 200 different files were merged into four tables namely, Cooling Degree Day (CDD), Heating Degree Day (HDD), Precipitation (PCP) and Average Temperature (AVGT). Moreover, those four tables were merged into the cleaned zip code and service data from the sponsor.

4.3. Data Exploration and Visualization

The lead data provided, required in-depth exploration. Since an exact definition of a good quality lead was not provided, the initial research was started by considering the factors which might qualify a good lead. It was initially decided that the right approach would be to focus on campaigns, source, and cost per lead because those are tangible factors that show where the lead is coming from and how much Treehouse was able to charge for that lead. When we requested more information and clarification from the sponsor about these factors, we were instructed that for the purpose of our analysis, both campaign and cost could be ignored. This resulted in having to dive further into both company and external data and it was decided that the revised approach would be to focus more on services provided by the associations per region, and exploring how external factors were connected or influential to the business.

External data, such as precipitation, temperature, soil quality, and groundwater have been explored, as these can be considered to potentially influence the market demand and therefore the leads for the different services associated with Treehouse.

There were a great deal of ways to approach the data so the analysis was started by visualizing the data in any way that made sense, before a coherent story about the business could be put together.

4.4. Association Visualizations

The analysis was started with a very broad overview look of all the associations and how they work together to create the Basement Systems brand (Slide 3). This showed the progression of the associations and their overall growth over the last ten years. Then a more detailed analysis was done to see how the leads that are distributed amongst these associations. When creating these visualizations, the time frame was set from 2011 given that prior to that time, there were

few active associations other than Basement Systems (Slide 4). There were also a variety of visualizations created in regards to leads coming in by region and state, so that it could be determined, where each association was most requested.

4.5. Service Visualizations

During the service level analysis, multitude of problems were encountered. This was mostly due to the fact that there are many services conducted by each association. Moreover, associations provide overlapping services. To remedy this situation, only the top ten most requested services were focused on for further analysis. But while performing the analysis and visualizations, some roadblocks were faced and a loss of information because of the restrictions put on the number of services. Another large issue with this was that Basement Waterproofing through Basement Systems is that largest service throughout all the Basement Systems associations, with a very high percentage of the total number of leads. Therefore, when only calculating information for the top ten services, the amount of leads for the other 9 services became almost negligible. When considering all the services we see value in many more of the work provided by the other associations.

Then a different mode of analysis was tried, by consolidating some of the service types through grouping in Tableau. Similar services were grouped together, including but not limited to:

- Roofing services
- Commercial services
- Doors and windows
- Fencing
- Basement

- Patios and Decks

It was assumed that these groupings would help understand the service data better however a similar problem arose as before. Grouping the services also created a large gap between basement services and others. Similarly, there are a lot of associations which handle certain basement services and not others such as basement waterproofing versus basement finishing. This created a lot of confusion when trying to visualize data by association. For this reason, all services were taken into consideration and they were analyzed by association.

When considering the growth of associations seen during the association level analysis, mostly Basement Systems, CN- Fence Nation, and CN - Roof Nation were focused on, given their growth trends (Slide 4). The services provided by these three associations and their locations were taken into consideration. This gave an idea about where the associations started and how they have grown overtime. It was also considered important, to look at the demographics that are found to most prominently request services from these associations (slides with bar graphs). While visualizing this data related to fencing, there were a lot of interesting factors which were observed. So, the most important variables to influence the demand for the service, could be easily found out. To better understand the factors influencing roofing needs, a few models were run to characterize importance in the variables.

4.6. Models

The prediction model was performed on data consisting of roofing service information from 2016. The target variable was the number of leads per zip code. A new independent variable named “Number of company” was created by counting the distinct dealers with each state. Then the data was split into training and validation with a ratio of 70%:30%. Population was removed to reduce multicollinearity.

Once the data was organized, a regression model was run in SAS JMP using Standard Least Squares method. Since the scale of x variables vary drastically, standardization of the variables was performed to address the issue. From the modeling results, female_population and male_population were found not to be significant in terms of p-value, so they were removed. Persons_per_house was also eliminated because it reduced the predictive power of the model. R-Square for training and validation is 0.5931 and 0.5422 respectively. Therefore, we don't have to worry about the problem of overfitting.

Since a regression model can only capture a linear relationship between the predictive variable and target variable, we built a boosted tree model with a non-linear relationship to supplement this. But we didn't apply standardization to this model because tree-based models won't be bothered by a variable scale. The input variables were the same in the regression model after variable selection. The number of layers was set to 50. Split per tree is 3 and learning rating is 0.1. R-Square is 0.6333 for training and 0.5921 for validation. There were no issues with overfitting.

Finally, we compared the regression effect summary from the regression model with results from the boosted tree model. Four variables were common and predominant in both models: Number of company, income_per_household, precipitation and household_per_zipcode.

4.7. Visualizing Billable Rates

To understand how Treehouse is making money through its billable leads we looked at the trends of billable and no charge leads (Slide 13). Within Tableau, we looked at many of the services that are provided by the distributors but mainly focused on Basement Waterproofing, Fencing, and Roofing.

4.8. Sentiment Analysis

We utilized web scraping techniques on python and collected reviews from various websites such as Yelp, Homeadvisor, Consumer affairs, Basement Systems USA and Bird Eye. We took that data and aggregated into a corpus to get a sense of how customers perceive the services offered by Basement Systems. We then used RStudio to perform the basic steps of data standardization which included removing punctuations, cases, white spaces and stop words. We then put the corpus for stemming to extract the root word for all the reviews. NRC sentiment was then used to extract the sentiments for all the root words and the sentiment was binded to the corresponding word. The packages used for the sentiment analysis are:

- SNOWBALLC - It helps in implementation of Potter's word stemming algorithm for collapsing words to a common root to aid comparison
- TM - Allows for corpus handling, preprocessing and creation of term-document matrices
- WORDCLOUD - Lets users plot a cloud of words shared across all the documents in the corpus
- SYUZHET - This package allows for binding words to their respective sentiments. It utilizes four sentiment dictionaries to map the words in a robust way
- GGPLOT2 - Lets users create a variety of plots

Utilizing the above-mentioned packages allowed us to evaluate the body of text containing all the reviews and extract their sentiments across eight different emotions (anger, anticipation, disgust, fear, joy, sadness, surprise and trust) and two sentiments (positive and

negative). Scores were then computed across all the emotions and sentiments. The sentiments were later clubbed into more broad categories as shown below to aid the analysis:

- Positive - Positive, trust and joy
- Moderate - Anticipation, surprise and fear
- Negative - Negative, disgust, anger and sadness

5. Results

5.1. Analysis of Basement Systems

With the initial analysis it was found that Basement Systems is the most predominant of the associations with the largest number of leads. It has been around the longest and has had a steady stream of leads coming in on a regular basis. It was observed that the leads data for Basement Systems exhibited a prominent seasonal pattern. Quarter two (Q2) and quarter three (Q3) are the highest lead months for Basement Systems and that seems pretty standard. A noticeable point was, a large increase in Q2 of 2013 and a drastic drop in Q2 of 2016 compared to their previous years (Slide 4). To understand what may have caused this, the analysis was narrowed down only to the services offered by Basement Systems. Many services increased in Q2 of 2013 but the largest growth in the leads came from Basement Waterproofing, which increased by 111% (Slide 5). A negative change in the same service is what happened in Q2 of 2016, compared to the year prior, with a decrease in 18%. Next, a number of external factors were analyzed which could potentially affect Basement Systems services, especially Waterproofing and a correlation with precipitation was identified. Q2 of 2013 was much more wet than 2012 which may have caused the increase in basement waterproofing leads. A drought

in Q2 of 2016 is believed to have caused the lack of leads because little damage was done to basements across the USA.

5.2. Analysis of Growing Associations

When considering the smaller associations, it was found that many have grown over the last few years. CN-Fencing and CN-Roofing were the ones that particularly stood out with a drastic increase within the last two years.

5.3. Fencing

Apart from CN-Fence Nation very few records were found for other associations who did fencing work, therefore for the fencing service, CN-Fence Nation was focused on. While exploring the data provided, it was seen that fencing as a service is provided on mainly the East Coast, but has grown a lot in that area over the last few years (Slide 6). To analyze why there was such a dramatic increase in leads for fencing over the past two to three years, crime was considered. It was predicted that a rise in crimes may have caused a desire to protect their property with a fence. However, after looking into the crime rates in the United States over the past 10 years, we found that crime has actually gone down. Therefore, no direct correlation was found between fencing and crime rate.

It was found out that many more contractors were hired through CN-Fence Nation and perhaps the simple availability of the service in certain regions may be what has caused the increase in leads. Later it was confirmed by the sponsor that the number of distributors in the last two years have increased greatly. With the absence of sufficient and reliable data about CN - Fence Nation's competitors, it could not be confirmed if fencing as a service had already existed in the regions that showed such a dramatic increase in demand over the past few years.

Therefore, it was assumed that there was little to no supply of fencing services in the growing regions.

Other than an increase in distributors, some other factors were explored that could potentially influence the growing demand for fencing by looking at the patterns between the regions that requested fencing services. After some analysis, some specific demographics and geographical factors were found that were more likely to request fencing. Demographics include high income per household and houses with a high property value. Similarly, fencing demand was largely concentrated in regions with low elevation (Slide 7). High income and home value are expected characteristics of households with fencing needs given that fencing is the most luxurious service requested compared to the others provided by Basement Systems associations. Luxurious meaning it is not a requirement to constitute a livable space as basement waterproofing or roof repair would be. After our last discussion with the sponsor, elevation might not be the most informative characteristic due to the elevation levels of the east coast, where CN - Fencing is currently located.

5.4. Roofing

A similar analysis was conducted for CN - Roof Nation. Visualizations showed that CN - Roof Nation was the fastest growing "new" association (Slide 4). While exploring how it has expanded, we found that it is very similar to fencing, starting off in the New England area and expanding over many parts of the east coast (Slide 9).

The boosted tree model and regression mode were conducted to further understand the important demographics of roofing leads. These models found four variables that overlapped between the two models and were of highest importance: number of companies, income per household, household per zip code, and precipitation increment (Slide 10). Naturally, the

increase in distributors available in an area would increase the amount of leads in comparison to years in which distributors were not present. That being said, we did not use number of companies as one of our variable as one of our indicating factors. We focused on the other three and found that roofing leads come from areas with high precipitation, lots of households per zip code, and where the average income per household was relatively high around $70,000 per year (Slide 11).

5.5. Billable Leads

When considering the market mix of the associations, it was found to be shrinking for Basement Systems. They had started with nearly 100% of the contributing business sharing slightly with Total Basement Finishing. Now Basement Systems only accounts for approximately 30% of the brand's service leads. This is expected, given that many more associations have been added in the last ten years (Slide 3). The market mix for Basement Systems was decreased to make room for the other associations and their services. There is an increase in many services which affects the billable rates with each association.

It is also assumed that when an association is new, there is a lot of potential for Treehouse to get billable leads because the referral process has not yet taken effect. The largest association, Basement Systems, started as a high billable lead generator with many of its services, but certainly with their biggest money maker: Basement Waterproofing. The same goes for fencing which is mainly provided by CN - Fence Nation (Slide 13). As time goes on those billable rates decrease and this can be seen for almost all of the services. There are also occasions where a service goes from a low billable rate to a higher one but ultimately the rate will go below 50%. The increase in non-billable leads can be attributed to two major components:

- Distributors run television, radio, and other campaigns on their own.

- Associations and distributors are getting prominent through word-of-mouth, social media reviews, and other referral programs.

5.6. Sentiment Analysis

Positive, trust and joy were the top sentiments in the reviews, with over 45% of the words in the reviews being positive and trust based (A.1). This was a very good indication of what and how customers perceive Basement Systems as a whole and its associations and services. This also indicated that this huge positive and trust based perception could be turned into referral networks for increases business (Slide 14).

The word cloud generated also showed some positive yet interesting results (A.2). The most frequent words that customers used to express their opinion excluding the filler words and core words such as Basement Systems, work, job etc. include

- Great
- Done
- Happy
- Professional
- Recommend

Words such as professional, recommend and other positive opinions helped us validate the finding from sentiment analysis and led us to believe that all these customer opinions could be translated into creating referral networks for added business.

Another interesting observation was that the number of reviews grew tremendously in the recent years and we believe it will continue to do so as the customers are more tech savvy these days and hence are more likely to express their opinion on social forums. Additionally, the

expansion of various services has been very significant especially in the last 3 years and this will also generate more reviews and opinions.

6. Recommendations

6.1. Basement Systems and Seasonal Campaigns

We recommend that Treehouse considers seasonality when creating marketing campaigns for Basement Systems. Since we found that Q2 and Q3 are generally very demand- heavy in regards to waterproofing leads we can utilize Q1 and Q4 to inform prospects of what is to come. This could both increase awareness and increase demand from new leads as well as reduce the need for resources by moving Q2 and Q3 leads to the other questers. As we saw in precipitation, Q4 has been the rainiest quester for all the analyzed years, this means that Q4 campaigns could make potential customers aware of what could happen to their basement at any time due to the current wet season. These types of precautionary campaigns can inform the audience about what will happen if they do not waterproof their basement and the costs that will be associated if they are not proactive. Q1 marketing campaigns could target those that have been affected by the previous rainy Q4 season. According to our analysis there have been some rainy Q2 seasons, therefore, some Q4 campaigns can also be reused as a precautionary message. Before launching proactive campaigns to consumers in Q1, it would be necessary to check future weather forecasts to assure that it will in fact be rainy or will there be a drought again.

6.2. Expand Fencing

There are many factors that were shown to be consistent with leads that request fencing. Since this service is only available on the east coast, the demographics and elevation of those regions which show more demand for fencing on the east coast were analyzed and the regions on

the west coast were narrowed down on the basis of that set to demonstrate where CN - Fence Nation can expand to on the West Coast. Using the household income ranging from $58,000 to $91,000 per year, house value ranging from about $230,000 to $490,000, and average elevation ranging from 0 to 2,300 feet, we found that California and Washington state may be high-demand areas for fencing services (Slide 8). Similarly, Washington D.C. looks like it can benefit from having fencing services. Washington D.C. may be an area where it would be much easier to go forth with hiring more fencing distributors than the west coast given that there are multiple surrounding states with this service. Since the expansion of Contractor Nation in the east coast has proven to be successful in the last two years, considering tying up with more contractors and distributors on the west coast, in the regions that were previously stated, could help increase the number of leads in that area. Simply launching a trial period with a few contractors could help Contractor Nation analyze their return on investment and see if this venture is worthwhile.

6.3. Expand Roofing Services

With our results, we found that there may be demand for roofing in the west and central areas of the United States. Specifically, Texas, Colorado and Washington were states that shows to have characteristics similar to those found in the current incoming leads (Slide 12). To come to this result, we looked at an approximate precipitation increment range of .6 to 6.5 inches, approximate households per zip code range from 10,000 to 16,700 households, and a household income ranging from $62,000 to $72,000 per year. As suggested with CN - Fence Nation, it may be beneficial to first contract a few roofing service providers to come under the Basement Systems brand before making a full rollout in those states.

6.4. Focus on Services with High Billable Rate

We recommend that based on the natural life cycle of word of mouth Treehouse Group should focus a majority of its resources and energy on creating campaigns for services that have a high billable rate. We believe a billable rate of 60% or higher is a sufficient amount. This would ensure that Treehouse is spending its resources in the place where the return would be maximum. This is not to say that Treehouse should ignore associations that provide services which have a billable rate under 60%, merely to focus more energy on services which have not been so widely spread through word-of-mouth and referrals.

6.5. Invest in a new line of business

We recommend Treehouse to start a new line of business to moderate the social forums / review pages of distributors to create a new revenue and increase its revenue. Potential moderation responsibilities could include:

- Directing customers to dealer's websites through hyperlinks
- Addressing customers with negative opinion and helping in resolving them by getting in touch with them more privately to give a sense that Basement Systems is hearing its customers
- Sending out thank you notes to people who had a positive experience and potentially converting that customer into a referral for repeat business or added business from customers family and friends

We also recommend Treehouse to encourage customers to review after availing their services and send out surveys asking for customer opinion after availing services. This would ensure Treehouse the following:

- More reliable and large volume of data to interpret the customer opinion

- Create a feedback mechanism to improve its services and efficiency
- Help in creating a referral network for added and repeat business

7. Conclusion

As a business which has been successfully bringing its brand into the homes of United States citizens, there is always room for improvement. There is a lot of potential for Basement Systems to expand into other part of the United States where regions are untapped with Basement System's services. However, further research is certainly required. The most important information that must be collected before any expansion decisions are established is a competitor analysis. It is vital to know how many other companies are providing services such as waterproofing, fencing, and roofing in other parts of the United States and how saturated that market is already. This will allow Basement Systems and Treehouse to understand if certain areas are in fact profitable.

8. References - Data Gathering

(2015). *House Prices: Advanced Regression Techniques.* Retrieved from https://www.kaggle.com/c/house-prices-advanced-regression-techniques/data

(2017). *Precipitation Data Summary*. Retrieved from http://ftp.cpc.ncep.noaa.gov/htdocs/products/analysis_monitoring/cdus/prcp_temp_tables/mctyprcp.txt

Angie's List Reviews. Retrieved from https://www.angieslist.com/companylist/us/md/elkton/basement-systems-usa-reviews-2327313.htm

Basement Systems USA Reviews. Retrieved from https://www.basementsystemsusa.com/about-us/reviews.html

Home Advisor Reviews. Retrieved from https://www.homeadvisor.com/rated.BasementSystemsUSAInc.46223166.html

Leader Basement Systems Reviews. Retrieved from https://www.leaderbasementsystems.com/about-us/reviews.html

National Centers for Environmental Information. (2017). *Climate Data Online: Dataset Discovery.* Retrieved from https://www.ncdc.noaa.gov/cdo-web/datasets

National Climatic Data Center. (2017). *Average Annual Precipitation by City in the United States.* Retrieved from https://www.currentresults.com/Weather/US/average-annual-precipitation-by-city.php

United States Census Bureau. (2017). *Housing Vacancies and Homeownership (CPS/HVS): 2017.* Retrieved from https://www.census.gov/housing/hvs/data/q217ind.html

United States Census Bureau. (2016). *Characteristics of New Single-Family Houses.* Retrieved from https://www.census.gov/construction/chars/completed.html

United States Department of Agriculture. (2017). *Daily Soil*. Retrieved from https://wcc.sc.egov.usda.gov/nwcc/rgrpt?report=daily_soil_7day_state

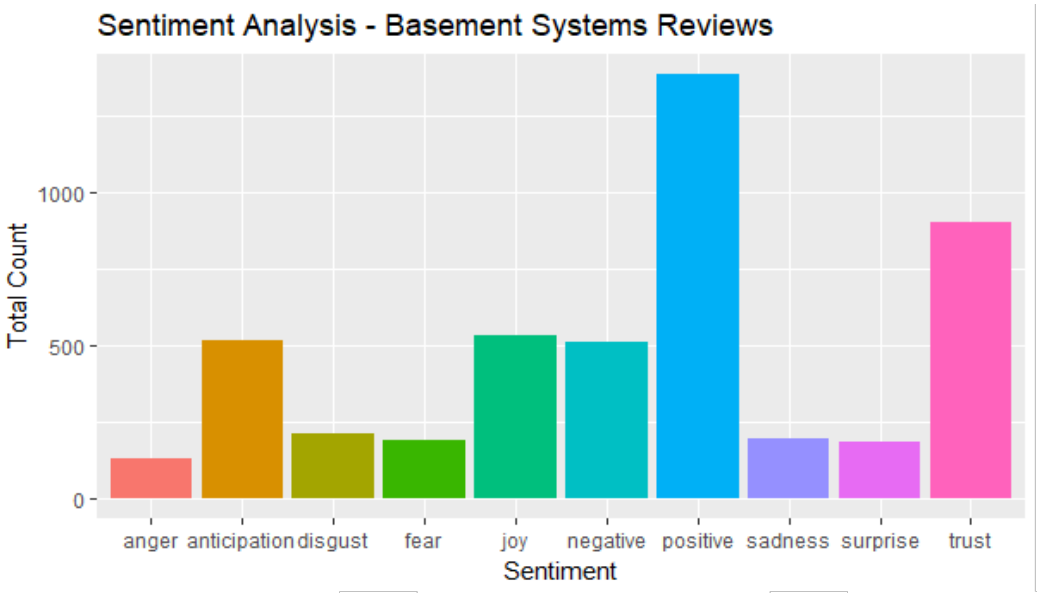
United States Geological Survey. (2017). *USGS Water Resources.* Retrieved from https://waterdata.usgs.gov/nwis/current/?type=gw

Yelp Reviews. Retrieved from https://www.yelp.com/biz/basement-systems-usa-elkton-3

9. Appendix

Sentiment Analysis

A.1 - Overall Sentiment



A.2 - Word Cloud