****

DSO 545 Statistical Computing and Data Visualization

Project Final Report

Professor Sharif Abbass

Cosimo Arnesano –arnesano@usc.edu

Sean McCormick – spmccorm@usc.edu

Hong Gao – honggao@usc.edu

Xuewen Hu – xuewenhu@usc.edu

Jiarui Ding – [jiaruidi@usc.edu](mailto:jiaruidi@usc.edu)

# Executive Summary

Our team completed a deep analysis into 15 months of historical 24Hr HomeCare sales data, and prepared several recommendations based on these analyses. Our hope is that management at 24Hr HomeCare will conduct further research into our conclusions and consider implementing these changes with the goal of increasing sales, reducing costs, thus, improving profitability. Our primary recommendation addresses the fundamental lack of caregivers that 24Hr HomeCare (and other homecare agencies) experience. Through analyzing the average distance that a caregiver travels from their home to the home of their client(s), we found a negatively correlated relationship between the average monthly value of a caregiver (how much revenue they earn for the company) and the average distance they travel to work. This theme was evident after looking at the address locations plotted on maps (see Torrance location on the Dashboard for customer and caregiver clustering), but it was confirmed through visualizing average distance traveled weighted by average revenue earned by caregivers. We suggest that 24Hr HomeCare invest resources in 1. Matching caregivers to clients while being more cognizant of the distance between them and 2. Focus caregiver recruiting efforts in branch locations that have a higher average distance traveled and lower average caregiver revenue earned.

We have appreciated the opportunity to collect, clean, manipulate, and analyze 24Hr HomeCare’s historical data, and hope that leadership within the company takes action based on the insights and recommendations that we have provided here. Please see the remainder of the report for further detail on our methods, calculations, and to recreate our analyses using R.

### Introduction:

24Hr HomeCare is a company based in El Segundo, California that provides non-medical homecare services to the elderly community, injured workers, and the developmentally disabled. They have granted access to the historical sales, customer, and caregiver data for the senior care business segment from Quarter 1 2017 through Quarter 1 2018 in the hopes that our group may derive some insights from analyzing this data using statistical techniques and visualization that have not been done before. The “baby boomer” population is entering retirement age, and a desire to retain their independence in retirement has enabled homecare agencies to grow in recent years.

The company is operated by 12 branch locations throughout California, Arizona, and Texas. These locations may have unique properties, and should be analyzed individually. The business is broken up into three segments; services provided to the elderly, services provided to injured employees (funded by worker’s compensation insurance), and services provided to the developmentally disabled (funded by government grants). Due to the complicated nature of the disability services segment (resulting from government funding), our data focuses only on senior care and injured workers.

24Hr HomeCare hires caregivers in communities surrounding each branch location, and pays them an hourly wage to provide homecare services to the customers we match them with. We bill the customer an hourly rate that includes a markup (usually around 45%). Regions in California and the states of Texas and Arizona all have differing minimum wages, which drive up or down the wages that are paid to caregivers in those areas, and subsequently affect the bill rates that can be charged to those customers.

A primary factor limiting growth in the homecare industry is the supply of caregivers available to staff shifts with customers. Recently, the human resources department has inquired whether a correlation exists between the distance that a caregiver must travel from their home to the home of the customer and the tenure of that caregiver (i.e. if they must travel farther, will they stay on as a caregiver longer). In addition to the insights gained from analyzing distance, we will be doing a deep dive into other metrics that have the potential to drive the revenue and profitability of 24Hr HomeCare.

# Data Description:

### Navigating the R Files:

In order to recreate the analyses contained in this project, please load the project “homecare\_final” and then perform the following steps:

1. To generate the master data frame (“shifts”), open “cleaning.R” and run the entire script
2. To open the Shiny Dashboard, open “24HrHomeCareApp.R”, select the entire script then run everything (clicking “Run App” will not work)
3. Open “analysis.RMD”, and click knit to generate our analysis document

### Description:

The data was gathered from the homecare company’s data server using SQL, and was converted to CSV files so that they may be imported into R. The data includes the following tables and fields:

1. Customers (Clients) – Contains all historical customers from 2014 to present
   1. Unique Key – Used instead of Customer Name for HIPAA compliance purposes.
   2. Address Data
      1. Customer Street Address, Street Address 2, City, State, Zip Code, Country
   3. Branch Location – The home branch office that the caregiver belongs to
   4. Demographic Data – Gender, Birth Date
2. Caregivers (Employees) – Contains all historical caregivers from 2014 to present
   1. Unique Key – Used instead of Caregiver Name to protect their privacy.
   2. Address Data
      1. Caregiver Street Address, Street Address 2, City, State, Zip Code, Country
   3. Demographic Data
      1. Gender, Hire Date
3. Shifts Worked (Split into 1 and 2 due to Excel’s limit of 1.048M records) – Contains historical shift records from January 2014 to March 2018.
   1. Revenue – Total revenue earned from the shift
   2. Labor Cost – Total payroll paid for the shift
   3. Hourly Bill Rate – Bill rate that was charged to the customer
   4. Hourly Pay Rate – Rate that the caregiver was paid
   5. Billed Hours (Regular, Overtime, Double Overtime)
      1. Hours that were billed to the customer, may differ from hours paid to the caregiver
   6. Paid Hours (Regular, Overtime, Double Overtime)
      1. Hours that were paid to the caregiver, may differ from hours paid to the customer
   7. Shift Date – The date that the service was provided to the customer
4. Location
   1. Connects Location Key to Location Names
      1. i.e. “13” is the *Location Key* for the *Location Name* “Torrance”

# Data Cleaning:

As can be expected when dealing with real-world datasets, the data from 24Hr HomeCare required a significant amount of cleaning prior to generating meaningful analyses. This data was extracted directly from 24Hr’s data source using SQL queries, and converted to CSV files. Our original goal was to analyze 51 months of shift information, but we quickly found that the volume of records was difficult to manage, and instead decided to focus our efforts on shifts worked in 2017 and Q1 of 2018 (as this more recent data will be most useful in prescribing recommendations to the business).

1. Combining shift files
   1. As mentioned above, due to the more than 1 million shift records, separate CSV files had to be loaded into R, and then combined using the ***rbind*** function into one master Shift table.
2. Formatting Data Types
   1. **Dates:** All date fields that were loaded (hire date, shift date, birthday, etc.) were converted using the various formulas within the Lubridate package, as we intended to use the date dimension in many analyses and calculations.
   2. **Revenue/Labor:** The Revenue and Labor fields were recognized as Factors, which is not useful when trying to summarize. These were first converted from Factor to Character using the ***as.character*** function, and then to Numeric using the ***as.numeric*** function (as converting directly to Numeric was resulting in incorrect values).
   3. **Addresses:** Address fields were recognized as Factors, so to enable analysis of those fields, we were required to convert them to the Character data type.
3. Manipulations and Joins
   1. **Addresses:** As one of our main goals of this project was to determine if a relationship exists between the distance traveled by caregivers to the homes of their customers, we needed to concatenate our address data into a single field for both customers and caregivers. We achieved this by using the ***unite*** function of the Tidy package. We then joined this field to our shift table using Dplyr’s ***left\_join*** function. We chose the ***left\_join*** function because we did not want to bring over any customer records that did not exist on the shift table.
   2. **Locations:** Present in the Customer, Caregiver, and Shifts tables was a numerical “Location Key”, and the Text “Location Name” was stored with the key in the Location table. We used ***left\_join*** to make the more analysis friendly “Location Name” field available in the other tables.
   3. **Various Field Name Changes:** As much of our cleaning required joining tables together, we were required to adjust the primary and foreign key field names to be uniform. We accomplished this using the ***names*** and ***paste*** functions.
4. Data Filtering:
   1. **Customer Demographic Data:** We wanted to be able to make meaningful analyses using the customer’s birthday and gender, so we filtered our customer list to only include customers with valid birthdays and genders.
   2. **Addresses:** During the process of concatenating our address fields together, we determined that many customers and caregivers did not have complete or accurate address data, so through a combination of identifying clearly incorrect values and “NA” or “Unknown” values, we filtered both the customer and caregiver lists to only those with valid addresses. We understood that this cleaning would likely not catch all mistakes, but we knew that the ***geocoding*** process would identify the remainder of invalid addresses.
   3. **Revenue:** There are several situations that arise where a shift is entered but a customer is not billed for services (caregiver interview, training shifts, orientation, complimentary care, etc.), but for the purpose of our analysis, these shifts were not relevant to include, so we removed any shift where the Revenue amount was not greater than 0.
   4. **2017 and 2018 Shifts:** As mentioned above, for practicality purposes we chose to reduce our dataset to shifts that occurred on or after January 1st, 2017.
   5. **Final Shift Table:** After performing the cleaning mentioned above, we joined our Caregiver and Customer tables to our Shift table (using the primary keys “EmployeeKey” and “CustomerKey”), which resulted in a Shift table that contained all relevant fields. We did this because the Shift table contains the lowest level of detail, and is the best single table to use for performing our analyses and visualizations. Finally, due to the removal of customers and caregivers with incomplete address or demographic data, we had to further filter our Shifts table to only include shifts performed with both valid caregivers and customers.
5. Calculations:
   1. **Latitude/Longitude:** Our intention was to use the ***geocode*** function of the ggmap package to derive the latitude and longitude from the concatenated “full address” information. We quickly found that this function and package contained some serious limitations that prevented us from collecting the coordinates for our clients/caregivers. These limitations included the 2500/day query limit to Google’s servers, the inconsistency with which it would derive the coordinates (of 1000 addresses queried, it would successfully provide longitudes for 300 addresses and latitudes for a different set of 400 addresses, as an example), and finally, the query considered latitude one request and longitude another, so our 2500/day limit was effectively cut in half. This limitation of google was another contributing factor in our decision to reduce our Shift dataset from 51 months to 15 months. After speaking with the professor, we exported our final list of customer and caregiver addresses (with their respective keys, so that we could reconnect the values to the shift table). Unfortunately, most geocoding services require delineated address information, so we were required to update our Shift table to also include the individual address fields. We used the U.S. Census website [1] to generate the geocodes, and after validating some sample codes, we imported them and joined the fields to our Shifts table. As expected, some addresses did not generate geocodes, so the shifts related to those customers or caregivers were removed from the Shifts table. We were pleasantly surprised by the relatively low number of NAs provided.
   2. **Haversine Distance**: While having the coordinate data would allow for mapping of coordinates on map plots, we wanted to dig deeper to see if we could identify a relationship between the distances caregivers travel for their shifts and other metrics (namely, caregiver tenure or revenue earned). We did some research on different methods of calculating distance, and landed on the Haversine formula, which is unique due to its consideration of the curve of the earth. Using the ***distHaversine*** functionof the geosphere package, we were able to determine the distance between the caregiver and customer on each shift record in meters, which we then converted to miles. We used the ***group\_by*** and ***summarise*** functions of the Dplyr package to generate a list of average distances traveled by caregivers (EmployeeKey), which we then joined back to the Shift table. Based on an analysis using a histogram of shift distances, we chose to ignore shifts that had a distance greater than 50 miles, as these outliers were likely incorrect.
   3. **Customer and Caregiver Lifetime Value**: Similar to the method used above for determining the average distance traveled by caregivers to their clients, we used ***group\_by*** and ***summarise*** to find the total revenue earned by customer and by caregiver, and we then linked those calculations back to the Shifts table.
   4. **Customer Age:** Differencing today’s date and the birthday field, we were able to use the ***mutate*** function of the Dplyr package to calculate the age of our customers.
   5. **Length of Stay:** With the eventual goal of calculating each caregiver and customer Average Monthly Revenue, we needed to start with calculating their Length of Stay, which we accomplished using the ***group\_by*** and ***summarise*** Dplyr functions, differencing the days between the earliest and latest shifts in the dataset on a per customer basis. We then joined this difference back to the Shifts table.
   6. **Average Monthly Revenue:** This metric was important for us to calculate, as it gives a better indicator of the value of the customers and caregivers than their “Lifetime Value” when we are looking at a dataset of 15 months. This effectively normalizes the “benefit” attained from a customer that happened to be receiving care during all 15 months when compared to a customer that may have started care at a later point during this period. We calculated Average Monthly Revenue by dividing the Customer Lifetime Value by the Length of Stay, and then multiplying that value by 30 (as Length of Stay was calculated in days). As with most other metrics, this was accomplished by using Dplyr functions.

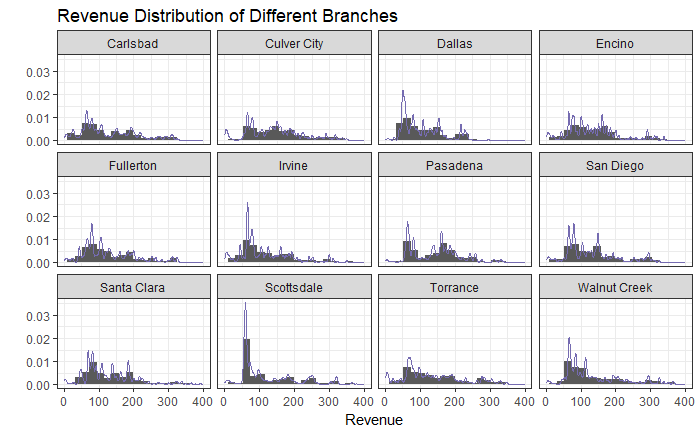
# Exploratory Data Analysis:

In this part, we take the “cleaned” data from last part and do some analysis and visualization with it. First, we applied a Twitter analysis on homecare industry to see what people are saying about this business. The word cloud shows that "jobs", "parents", "elderly", "help" are frequently mentioned on Twitter, which could mean that there's clearly a board labor demand in this market. In fact, a primary factor limiting growth in the homecare industry is the supply of caregivers available to staff shifts with clients, which is why this project will focus on analyzing the relationship between customers and caregivers.

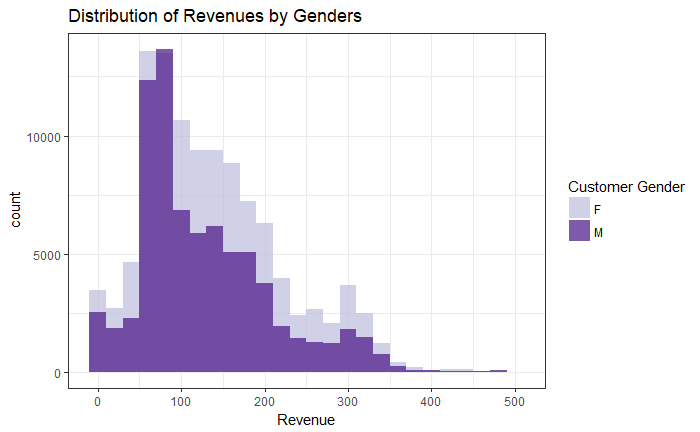


Since the outliers could affect our result, we focus on revenues under 200, which shows that branches have same kind of business emphasis or target group of clients in general, that makes most of firm’s revenues come from that range. The shapes and numbers of the peaks reflect that the main revenue sources clients of different branches/areas don’t have to come from only one group of clients or business. Some branches don’t have very obvious peak(s) but have their revenue sources distribute more evenly than other branches (e.i., Culver City, and Torrance). The reason could be that these branches are involving many and more business evenly and have the revenues come from all of these business sectors. However, they don’t have a focus or strength among all these business sectors.

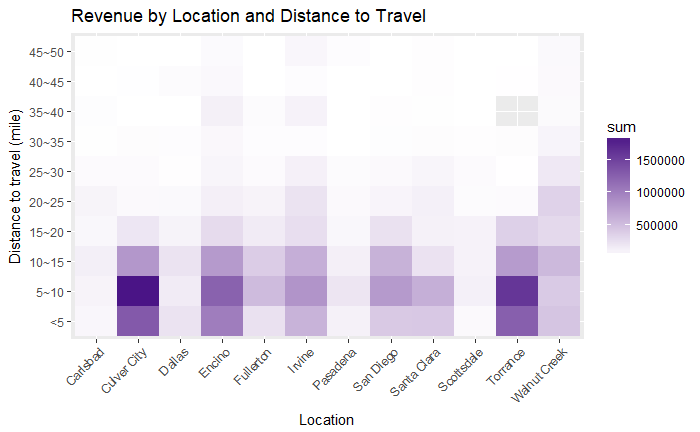
The branches with two peaks distributions (e.i., Pasadena, Dalla, San Diego, and Carlsbad) have their main revenue from two kinds of business and the revenue of these businesses are different. On the other hand, for the one-peak distributions, especially Scottsdale and Irvine who have very high peaks, they have more focused business, because same kind of business tend to bring the same range of revenue.



The distributions of Revenue by gender shows some kind of similarity around 100. For the revenue range $100-$300, female account more than male. In the meantime, the revenue peak for female is higher than male. Hence we can conclude that female spend more and there are more female customers. Obviously, company should pay more attention to female customers.



Distance to travel among cities doesn’t vary much. The most frequent travel distance is 25~30 miles, companied by the fewer distance to travel, the more revenue in total. And we can also tell that cities earn most are Torrance and Culver, while Dallas, Pasadena and Scottsdale end up with lowest revenue sum. But it can also be seen as them having more growth opportunities.

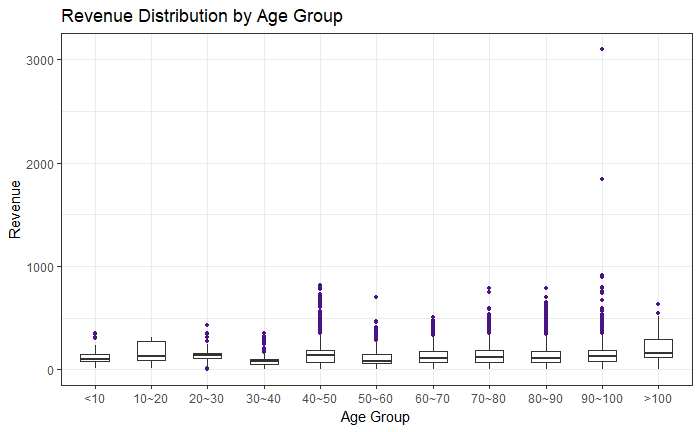


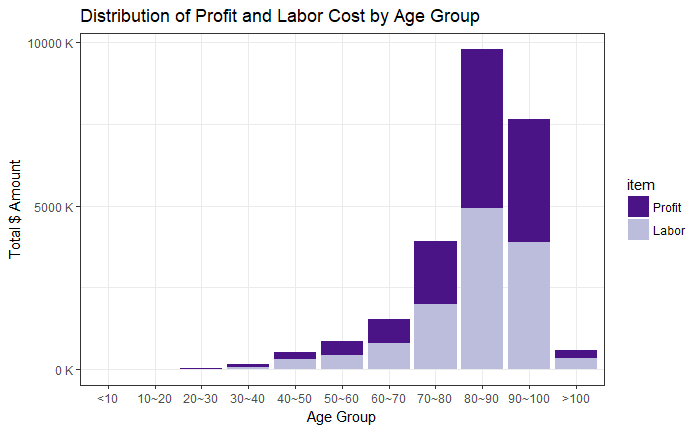
We build the following three graphs to analyze the relationship between customers’ age and the company’s operation results.

Customers aging between 80 and 100 composite the biggest group and bringing the most revenue among all customers. The median revenue of each service provided for customer over 100 is higher, as shown by the median line in boxplot, however, the high labor cost may make this group a less desirable target to expand business and increase profitability.

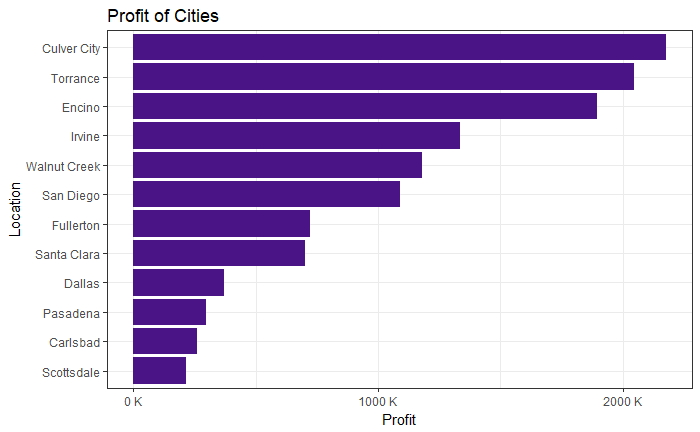
There might also be potential growth of profitability in customers younger than 40, however, considering that most of the services required are resulted from occasional accidents, for which unpredictability can be high, these groups might not raise main strategic attention.

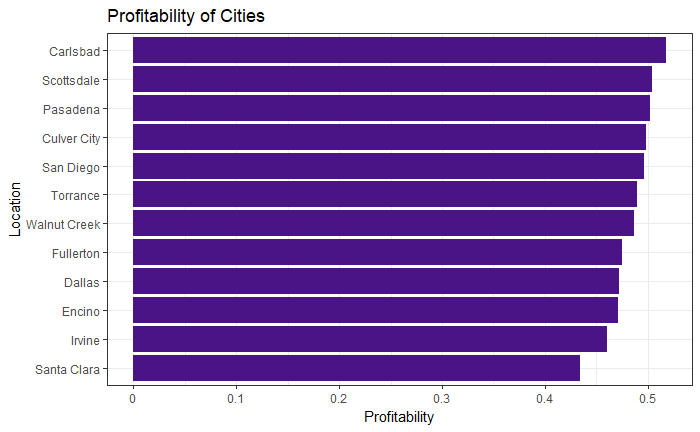
For the rest groups, consisting of customers aging between 40 and 80, the performance of profit and revenue are both neutral. According to the boxplot, outliers with high revenue to the company play a significant role, just as customers aging over 80. Therefore, given that the services provided for these groups are much less than people aging over 80, and the possibility that there might be more new customers to explore, 24Hr HomeCare may profit from increasing the customer base in this area.





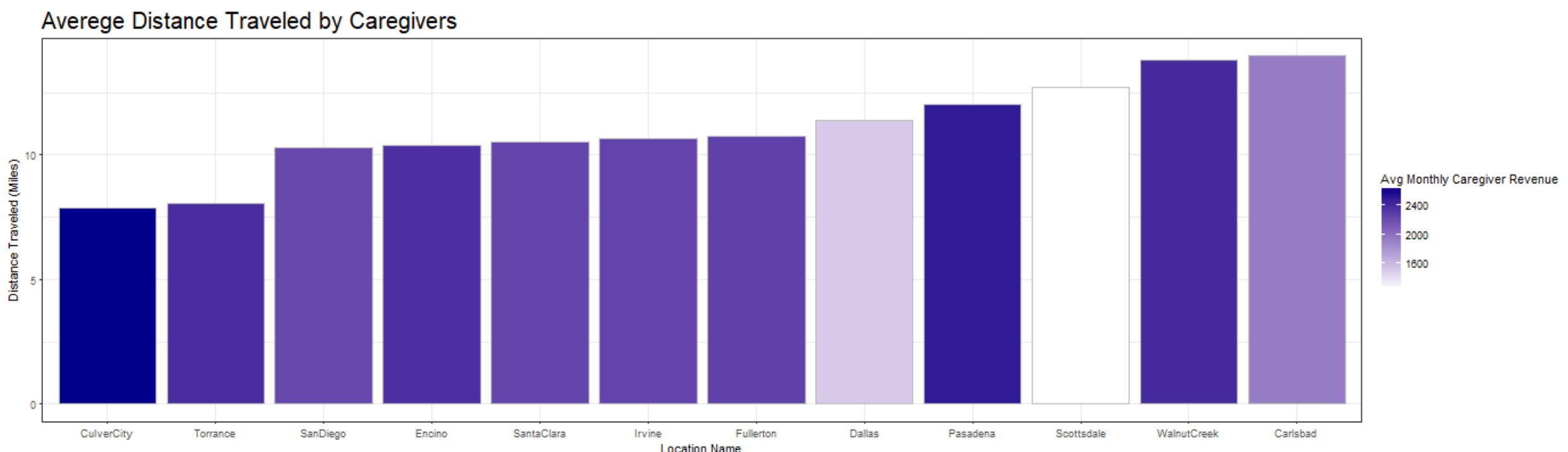
Profits of different cities varies significantly. Among them, Culver City, Torrance and Encino contribute the highest profit. But if we take a look at profitability, here we use Profit/Revenue to represent it, we can see that profitability ratios are pretty close. Carlsbad has the highest profitability and Santa Clara has the lowest among them. Since Carlsbad, Scottsdale and Pasadena has a relatively low total profit and high profitability, we think the company could put more resource on those locations. The Economies of Scale could help the company grab more profit in these regions.





In conclusion, we identify female is the largest group for the company to explore and increase profits. Customers aging between 70 and 90 currently provides the most profits, however, age groups of 40 to 70 might has the greatest potential for performance enhancement. It’s also important to put more resource on cities with high profitability but low customer exposure, like Carlsbad and Scottsdale.

### Caregiver Location Information:



Using the map visualizations enabled by our dashboard, we were able to derive some interesting insights into the spread of customers and caregivers in our various territories. For instance, Encino borders the Culver City location, and while Encino has no clients in the Culver City territory, it is evident that Encino is using a large number of caregivers that live in Culver City. This indicates that Encino may be having an issue staffing their own shifts, and must borrow Culver City’s caregivers.

Additionally, when looking at the Torrance map, the clients that Torrance serves tend to live near the coast (which makes sense, as the area is more affluent). However, almost all of the caregivers live further inland. This does not appear to present any issues for Torrance though, as their average caregiver travel distance is second lowest in the company.

Finally, when looking at the last plot on the Dashboard (Average Distance Traveled by Caregivers, colored by Average Revenue earned by Caregivers), we notice a general trend: as the distance that a caregiver travels to their client, we see that, generally, the average revenue that that caregiver earns is lower. To extrapolate on this thinking, we can say that caregivers who travel more are not earning as much revenue for their branch location. Some notable exceptions are Walnut Creek and Pasadena, where the caregivers travel a long distance. Walnut Creek is a large territory, and the Average Monthly Customer Revenue is very high for them (determined using the dashboard histogram of Average Monthly Customer Revenue). Conversely, Pasadena appears to be borrowing significant numbers of caregivers from the Encino territory, which is increasing their average distance traveled, but their relatively low number of clients in a very affluent part of Los Angeles appears to be driving up their Average Revenue Earned.

Despite these outliers, the theme persists that as caregivers travel more, they earn less for the company. Further research would is required to conclude anything definitively, but one possible explanation for this is that “high-traveling” caregivers are working fewer and shorter shifts due to dissatisfaction over long commutes.

# Dashboard:

## Description & Function:

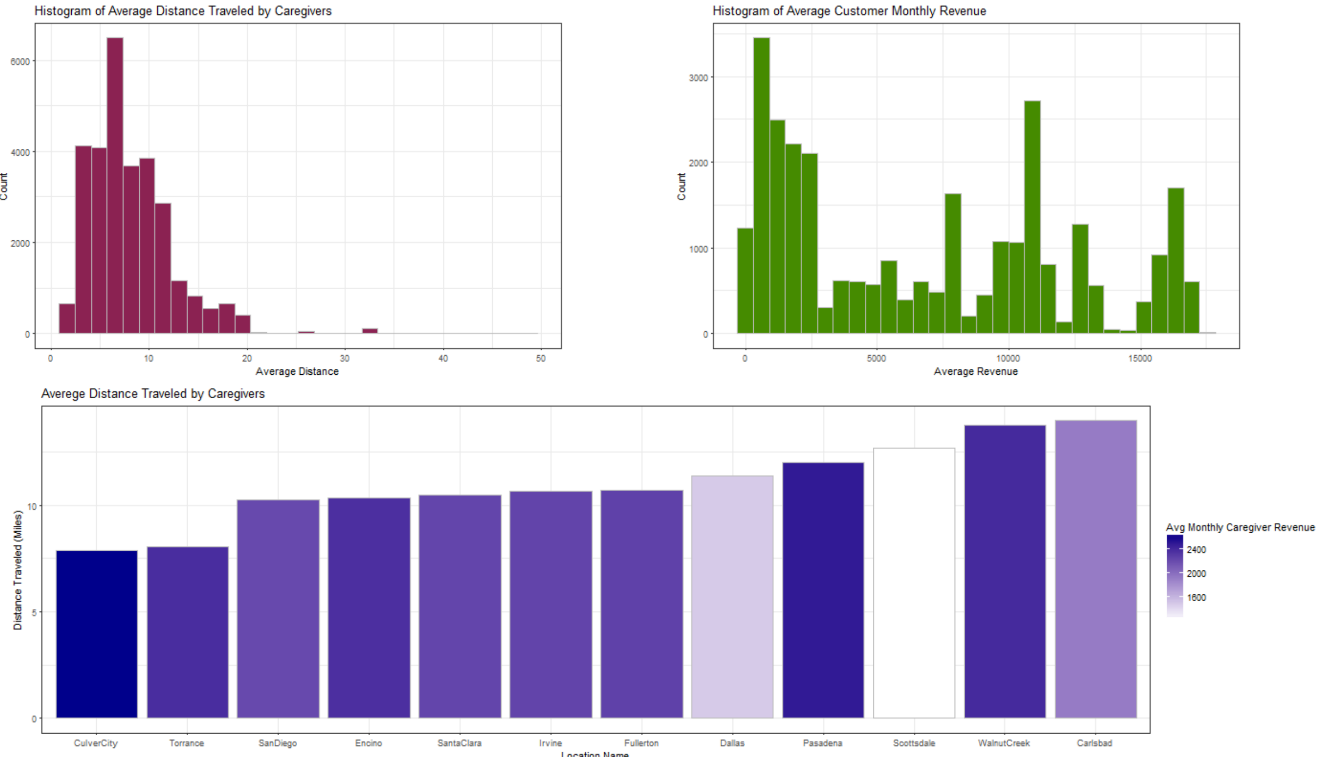
Our dashboard uses a drop down list that has the 12 branch locations available. Most of the graphs on the dashboard update when the location on this list is changed.

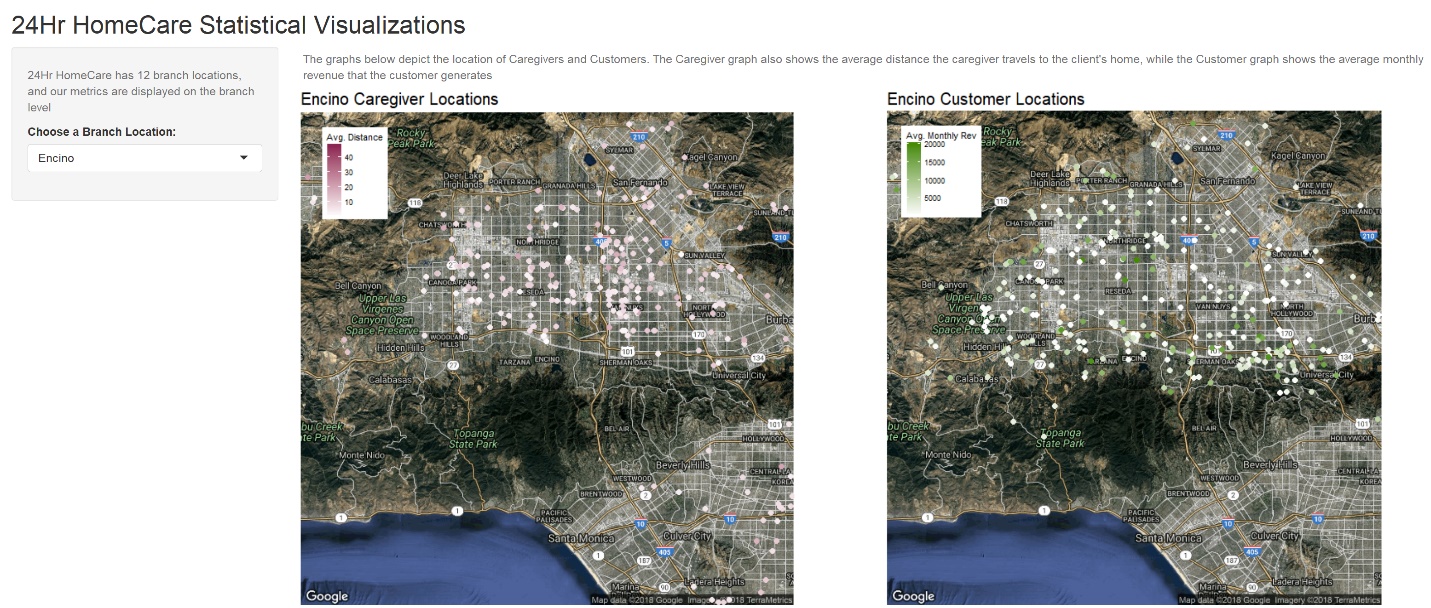
Our goal with this dashboard was to show each branch’s clients and caregivers plotted on a map of their own territory. We used the ggmap package, and the ***qmap*** function, to generate a map for each of the 12 locations. As with our previous experience, we found that most functions in ggmap were unreliable: they do not consistently load the map you are attempting to create. As with all of our coding, our goal is to have our scripts work 100% of the time they are ran. To correct for the instances where the ***qmap*** function fails, we used a retry formula [2] to create our own function that re-tries the ***qmap*** function upon failure, up to 20 times (while resting 1 second between each try). This enables our plots to run perfectly every time. Our dashboard includes dynamic chart names that update when the location is changed.

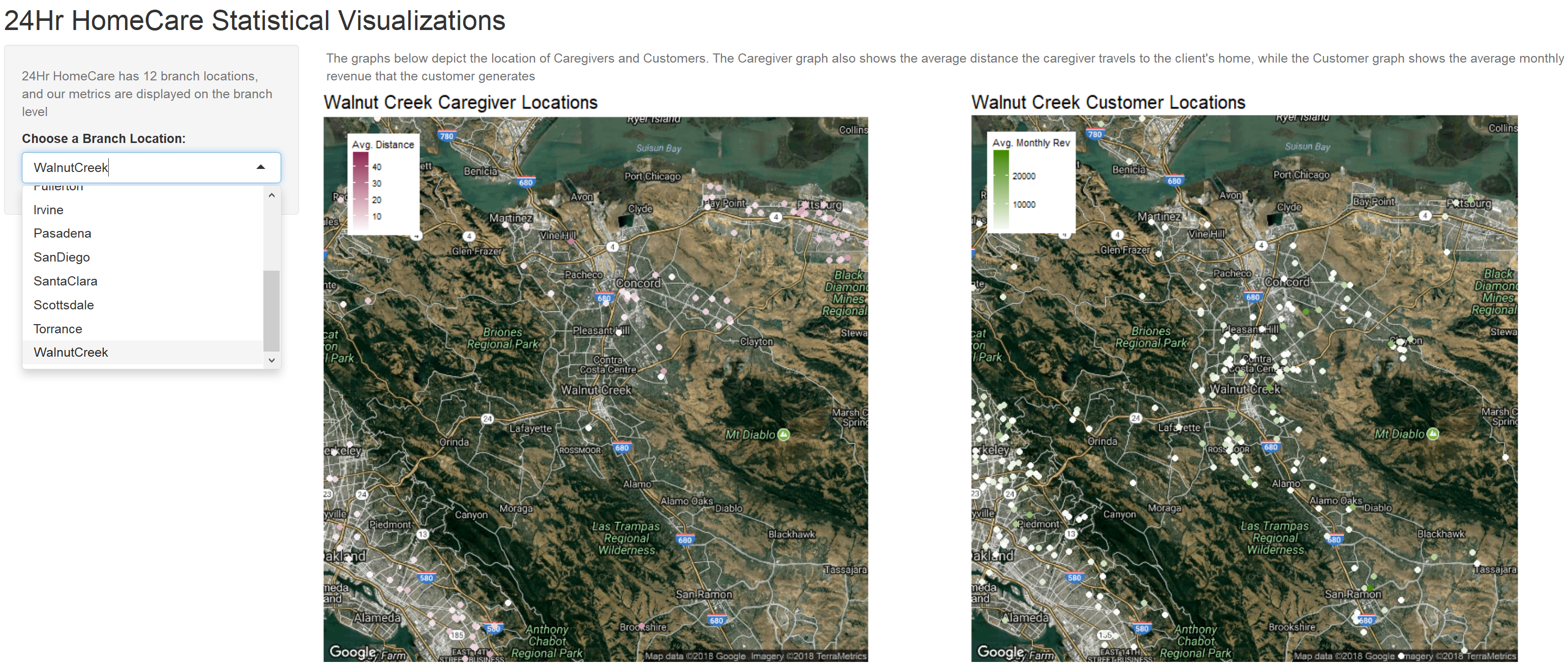
Using the latitudes and longitudes that we collected, we plotted the customers on one graph, and caregivers on another. We colored the dots based on insightful metrics (average distance traveled for the caregivers, and average monthly revenue for the clients).

Further, our dashboard also generates two histograms that give a deeper dive into the two metrics described above, so that the differences between locations can be viewed firsthand. Finally, the dashboard includes a bar chart that gets to the heart of the analysis: does the distance traveled by caregivers have an impact on the revenue that they generate on behalf of the company?

## Dashboard Screenshots:







# Conclusion/Future Work:

Data Integrity: As discussed, this analysis resulted in many business insights, but it also illuminated data integrity issues that 24Hr HomeCare should be cognizant of going forward. During our cleaning phase, we had to filter out a significant portion of customer and caregiver profiles due to incomplete addresses, missing demographic information, missing or incorrectly formatted birthdays, and other challenges. Stricter control of fields that are entered by users can help include as much data in analyses such as this going forward.

Caregiver Traveling Research: While it is evident that caregivers who travel more generally earn less revenue, further research with the HR department is required to determine whether this can be attributed to job dissatisfaction. If this assumption is true, and recruiting efforts can help convert “high-travelling” caregivers into “low-travelling” caregivers, then these caregivers may take on more shifts, take on higher length shifts, and potentially increase their tenure with 24Hr HomeCare.

# References:

1. Geocoding Census Data – Website: https://geocoding.geo.census.gov/geocoder/locations/addressbatch?form
2. Retry Formula - Website: https://stackoverflow.com/questions/20770497/how-to-retry-a-statement-on-error