Quality of Nursing Homes Throughout the U.S.

Identifying and analyzing factors that influence the quality of nursing homes and its correlations

Qiaoxue Liu, Walid Skeykho, and Christina Chen

CSE 163

Hunter Schafer

12 March 2021

PART 2

Summary of Research Questions and Results:

1. What is the correlation between the source of income and infrastructure (“Ownership Type”) for nursing homes in the US and the quality (as indicated “Overall Rating”) of its care?
    a. The data set provides the information about nursing homes, including what type of nursing home they are (their source of income and oversight) and tells us how well they are rated in their care. We can use the data of “Ownership Type” and “Overall Rating” averages for each to show the correlation between care quality and management type over the US.

    Results: From the analysis that we found, we found that in terms of averages, the non profit nursing homes were the ones that had the best overall ratings among the three. While the disparity between them is not very large, there is a clear correlation between the nursing home type and its review over the US. Shown on the geospatial graph of the US, non profit showed to have the most number of states with a very high overall rating, but we cannot assume the reasoning for this analysis.

2. Does more staff hours mean better care for the nursing home?
    a. Using the information of staffing hours per resident and different ratings for the nursing homes, we can show different distributions of how they correlate to better understand how important nursing care is to the overall health and success of these facilities. This is simply another distribution of how the quality of care at these nursing homes depends on a specific factor of that nursing home.

    Results: From multiple graphs that we generate, it shows that there is a slight correlation between the number of staffing hours and their overall rating. This correlation shows that if there is a higher number of nurses per resident, the nursing home is able to function better than if there are less nurses per resident that are available to help. However, we cannot make any assumptions of the reasoning for this trend but can only predict the overall result based off of the data.

3. Does a lower health quality rating have an effect on those living in the nursing home?
    a. An important aspect of medical facilities is the assurance that the patients and residents there are safe and healthy. Using reports of health care reviews for different cycles, we can track how much have the facilities improved over a three year span and track it along with the health deficiency scores to see how much of an impact it has made.

    Results: Based off of the graphs we generate, there does not seem to be any increase in the health ratings of the nursing homes. However, there is some overlap of the high density of higher rated nursing homes with the high health ratings of the nursing homes there.

4. Use geospatial data to find the density of high quality and best rated nursing homes in the nation and correlate that with the number of residents in nursing homes in the US. Does a higher demand mean higher rated care? (join countries dataset)
    a. We plan on presenting the overall data of nursing homes in the US on a geospasial graph and we will have it focus on two main details. This includes where the highest ranked nursing homes sit within the states and will be shown to present where the higher rated ones are most dense. We will also present the total number of reported residents in nursing homes to show where the demand is most needed to compare whether or not a higher demand for nursing homes mean that there will be more focus on improving the overall care.

    Results: As found from our data analysis and visualized with the maps, the state with the highest average overall rating is Alaska, indicating the location in the U.S. with the highest quality of care. The state with the highest percentage of nursing homes with rating 5 is also Alaska, supporting the conclusion that Alaska is where the highest rating nursing homes are most dense.

Motivation and background: This project is a small part of a larger context of analyzing medical care in the US. We want to be able to track the correlation between the ability for a nursing to care for its residents and the amount of service it can provide. Being able to understand the general trends of these nursing homes can help shed some light on some aspects of these facilities that should be focused on in order to help both the staff and the residents. I should be noted that we are aware of the common afflictions that come with assuming causation from correlation, but as mentioned before, these issues are not always restricted to a single characteristic. While nursing homes are not equivalent to hospitals or clinics (among other facilities), they all run under the umbrella of healthcare service, and can serve as another medium to analyze what actions these service providers have done so that they can do the best to help those who need it. Also understanding what types of facilities these are can also shed light onto what legislation might be lacking in this sector as to bring about a lower standard of care in comparison to other facilities.

Dataset:

➢ Information on currently active nursing homes: dataset name.csv

Description: This is a dataset from Medicare.gov provided by the Centers for Medicare & Medicaid Services. This dataset contains General information on currently active nursing homes in the U.S. in 2018. It includes the number of certified beds, quality measure scores, staffing and other information used in the Five-Star Rating System. There are 88 columns and 15340 rows, with each representing a nursing home.

➢ U.S. State Level Geospatial Data

This dataset contains the geometry data of the U.S. in terms of states, which is being used to create visualizations and maps of our findings.

➢ U.S. Zip Code Level Geospatial Data

This dataset contains the geometry data of the U.S in terms of zip codes, which is being used to create visualizations and maps of our findings.

Methodology (algorithm or analysis):

Our analysis consists of 2 stages: Machine Learning and Data Visualization.

Machine Learning

Preparation

a. We did data cleaning on the dataset so that it can be fed into the machine learning model for later analysis. We handled NA's by removing those rows, which did not remove too many rows from the dataset (700/50000).
b. Based on the types of the features, we did some transformations on it so that our machine learning model is more accurate.
    1. For continuous variables:
       We did log transformation on them because most of them are right skewed and there has been a big difference in the number.
    2. For categorical variables and years:
       We see if there are too many categories(more than 10). If there is, we combine some of them into a larger group to ensure that all categories have enough data points.
c. We also checked the correlation on the continuous variables to see if there are clear correlations between them.
d. Then we divided the data into 2 parts, a train set and a test set.

Train models

e. Since our response variable is "overall rating" which is a categorical variable, decision tree classification could be a good choice. We discussed in more detail the practicality of our assessment in the challenge section.

Result analysis

f. Analyze the result obtained from machine learning. See the accuracy of the model and make comments on whether the model is satisfactory. We also checked if our accuracy is better than random and ensure there is no major class classifier. Then explore some the influence of different hyperparameters (i.e. max depth of the tree) on the test accuracy and training accuracy.

Data visualization

g. We plan on using three data sets for the geospatial graphing. The first one is the data set about the nursing homes themselves, then there is a small data set that holds both the state names and the state initials so we can merge it with the third geo.json data set, and last, there is a geo.json data set that we will use to graph the nursing home information by states. Furthermore, there we also will be using plotly library data to show geospatial data of the nursing homes with an interactive display of the information. (Elaborated on in i. and j.)

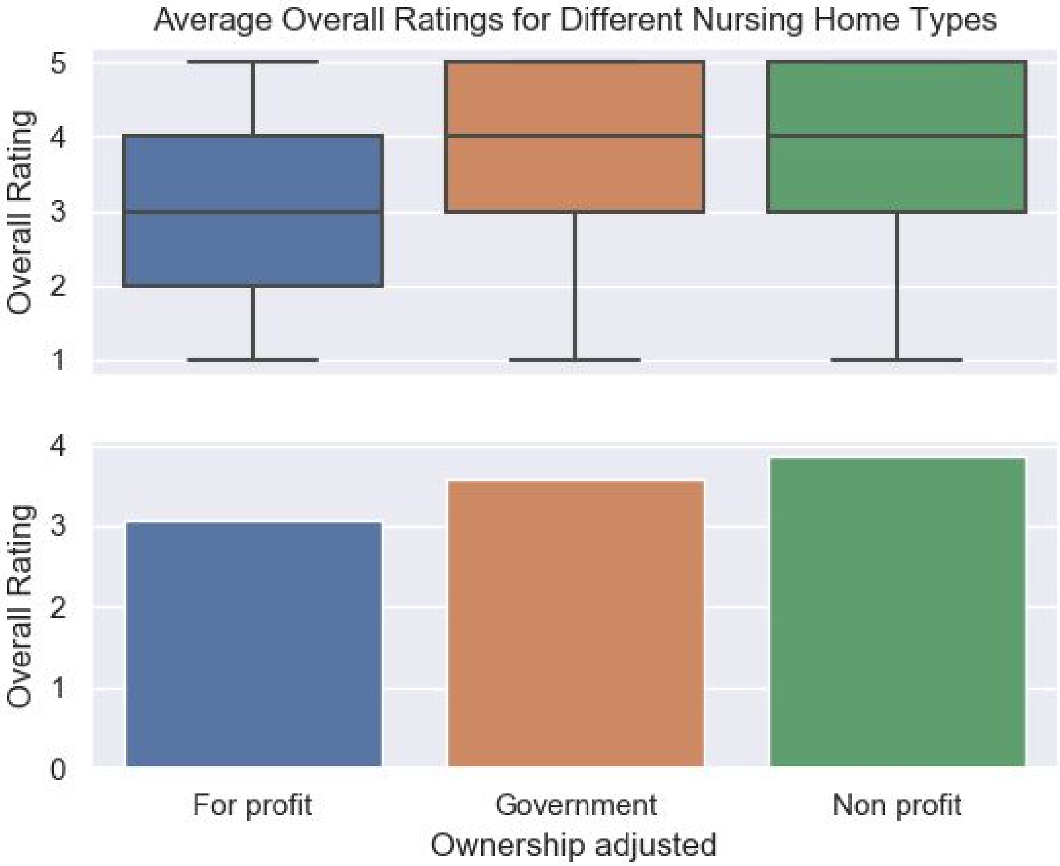
h. Another way we plan on displaying the data is through seaborn plots of line graphs, kernel density plots, bar plots, and scatter plots. Each of these graphs will be used to display a specific trend of information and will not be used twice on the same data.
i. Plot the nursing homes on the map and analyze the distribution.
    i. These maps are all in reference to either correlation plots that can present any significant disparities over the US based off of the data
    ii. Create interactive maps that can allow transitioning data
    iii. Make an interactive map, specifically a Mapbox Choropleth Map with plotly, to graph and display the average health care quality (aka. rating) for each state.
j. Using a new package called plotly, we were able to better visualize the data we analyzed with nice formats as well as interactive figures created. Using plotly's graph_object function we created visualizations of the overall average quality of care across the U.S. by each state. As well as the percentage of high quality care nursing homes within each state. Then using plotly's express function we were able to create an interactive scatter plot visualization the correlation between the quality of care and the number of residents present, with the trendline indicating the results.

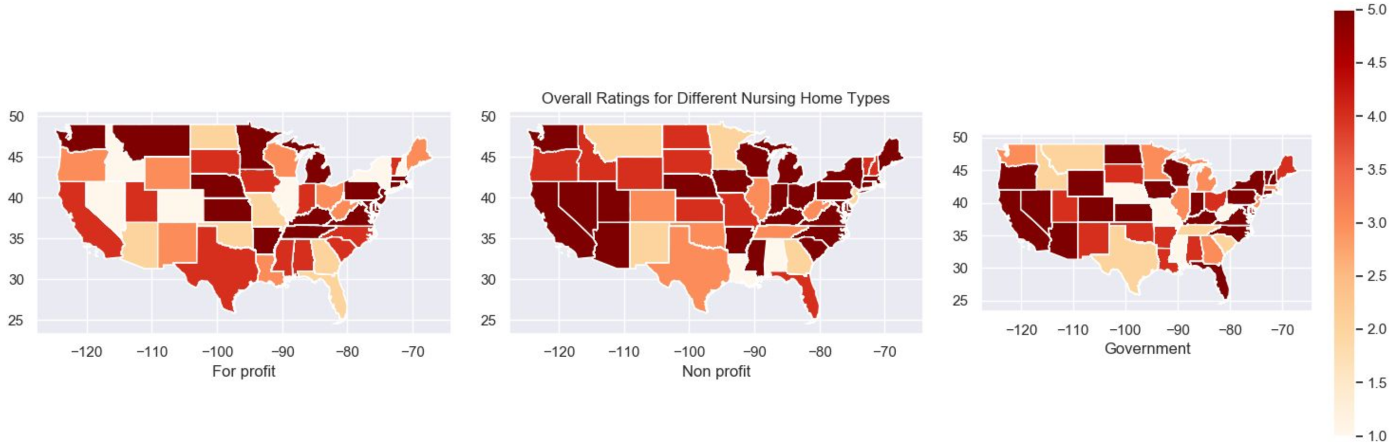
Results:

1. From these two graphs and a couple of individual calculations, it is clear that, in the US, there seems to be a higher overall rating of standard for non profit nursing homes in comparison to government run nursing homes and for profit nursing homes. While the information does not tell us the full picture of what might cause this trend, we can make a couple of assumptions and individual calculations. We found that in terms of averages, non profit also had the higher average reported nurse staffing hours per resident.

    This might be what aided the increase in efficiency and quality of care for these nursing homes, which would then lead to a higher health reporting and resident care reviews. In terms of locations, the geospatial data shows that the nonprofit and government nursing homes have similar results in terms of density of high rated nursing homes, but the profit nursing homes tend to do better in different regions. This could be due to the fact the nonprofit and government nursing homes probably run under similar jurisdictions that have set specific standards in those regions, while profit based facilities are capable of working under different infrastructures.

    But this does not mean that they are worse. Mostly because there are regions that the for profit facilities do do well in serving their residents. So it is mostly based off of that state and counties themselves and what laws and funding paths they provide for them.



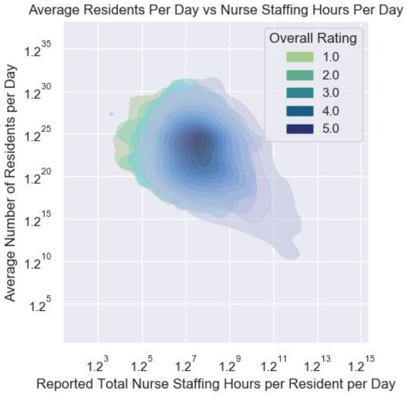


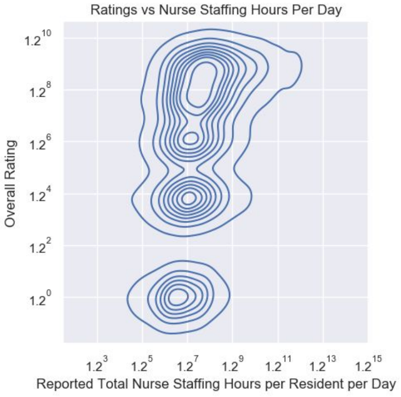
2. With this part of the analysis, we were trying to find what correlations would be found between the number of staffing hours, number of residents per day, and the overall rating of these nursing homes. With the first group of graphs and other individual calculations, we found that, with an increase in nurse staffing hours, there is a slightly higher probability of increasing the quality rating of the nursing homes. This is visually shown in the middle graph, where the higher

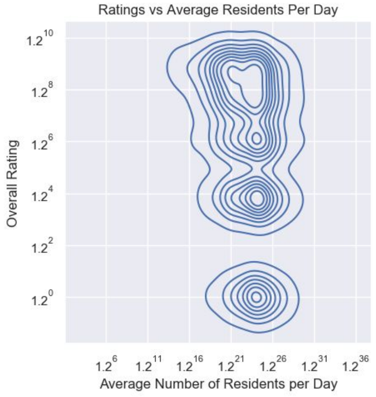
rated nursing homes curve toward an increasing number of staffing hours for the nurses. This is not surprising, and with a higher number of staff and supportive individuals in a facility, there are more people to help care for the residents and decrease the probability of any health related issues getting out of hand.

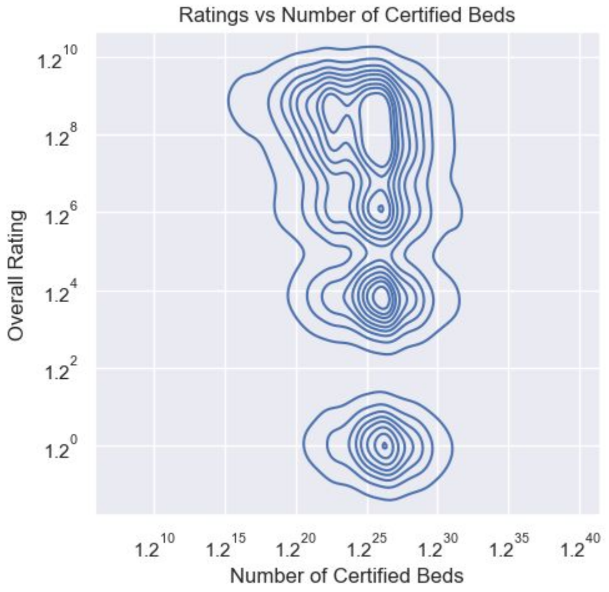
Another part that was shown on the third graph was that there was an increasing rate of higher recorded care when the average number of residents was less. This is not very dramatic, but there is a clear sign that with a decrease in the number of residents, there are less people that need to be cared for, so the quality of care increases for those individuals.

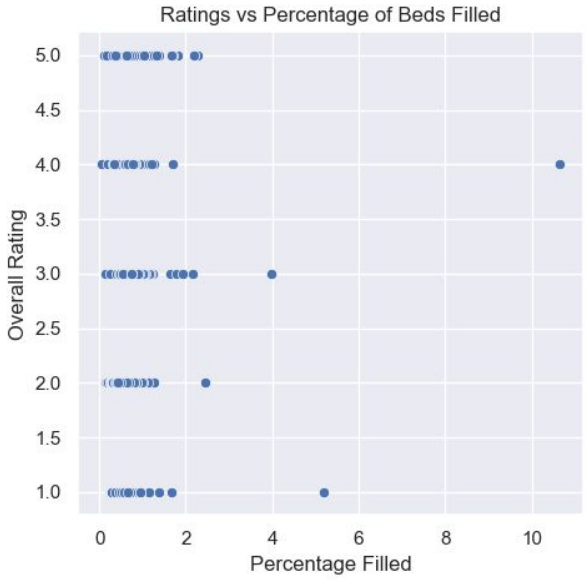
We found no extreme correlations between the number of residents in the facility and the average hours for staffing in the facility. We also did not find any correlations between the ratings and the number of certified beds. So, a larger facility does not mean a better one. The same goes for the percentage of beds that are filled up, this shows to have a very little effect on the quality of care. This data does not involve enough information to make a clear reasoning for this the lack of correlation.







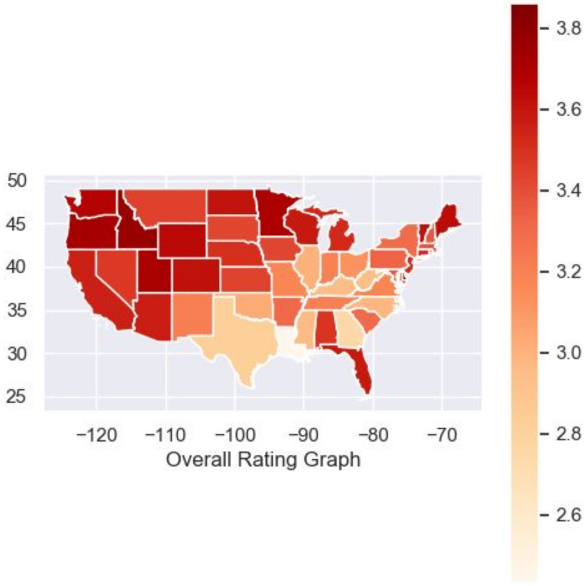


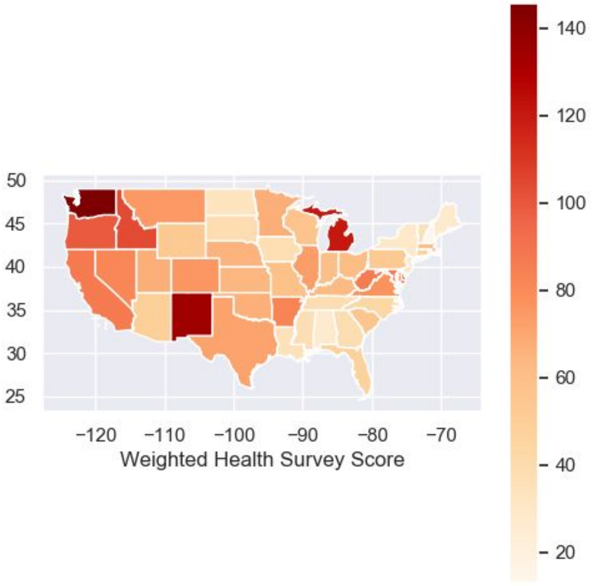


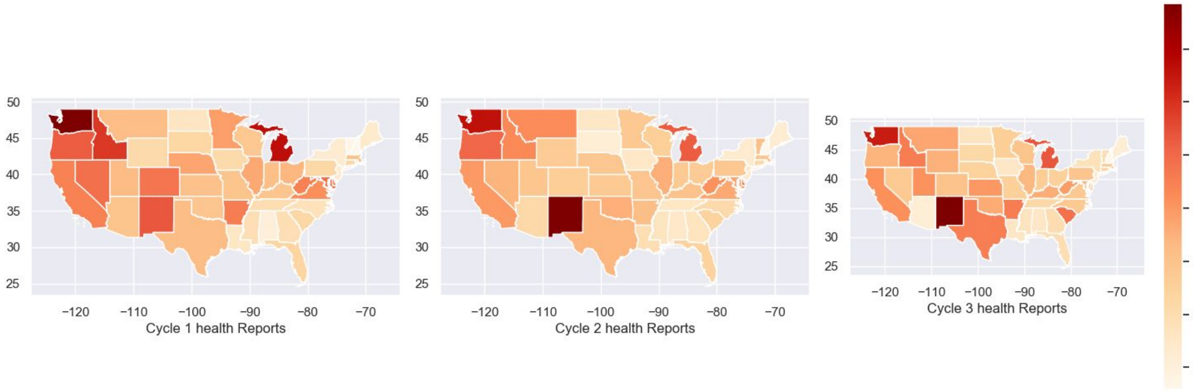
3. In this part, we tried to find whether there was a clear correlation between health ratings of these nursing homes and the overall rating of care for these nursing homes. We also wanted to find out how much these nursing homes improved their health ratings over their reported cycles. From what was calculated and shown, there is a slight correlation between the health rating and the quality of care that was reported from these facilities.

    However, I do not think that there is enough to make a rational assumption that if the health rating is poor in the facility, that means that it will not have a good quality of care. I thing that the higher health rating only improves the care, as it is necessary that a facility has good quality equipment and healthcare providers on hand for the residents. But, this is probably not what dictates the quality of the nursing home, as living conditions, nurses, facility activities, and even the location can help a nursing home take care of its residents.

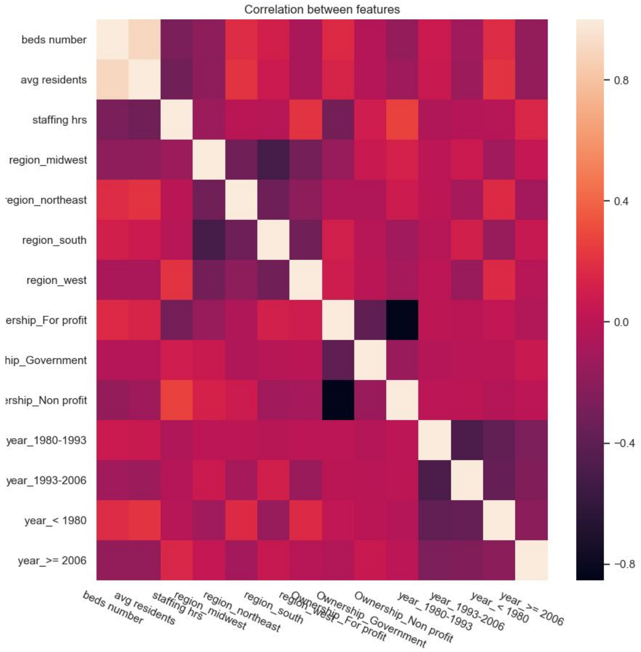
    The second group of graphs show the progression of the health rating for these nursing homes over a three year cycle. Surprisingly, it shows an overall decrease in the health rating of these facilities. While this is not true in all regions, it could be possible that there is some financial struggle during 2016 to 2019 that caused these facilities to depreciate in the quality of care, but this again is mostly speculation, and there are countless possible causes for the reported decrease in health care.



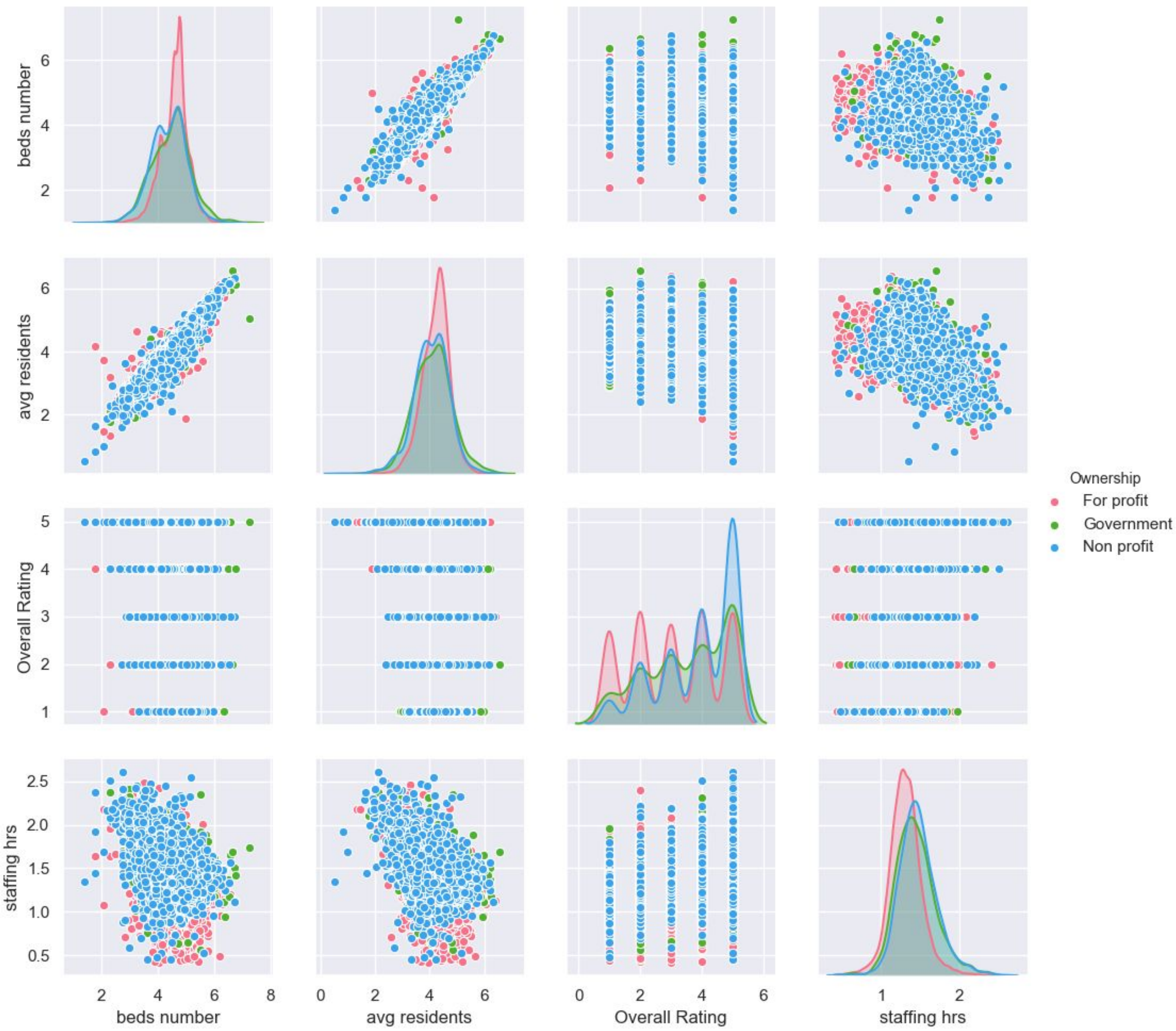




4. In the machine learning part, we found the following results:
    a. We show the correlation plots and see there is no clear collinearity between most of the features except for “beds numbers” and “average residents per day”. To improve our model in the future, I think we could remove one of the columns while keeping the other as the feature.

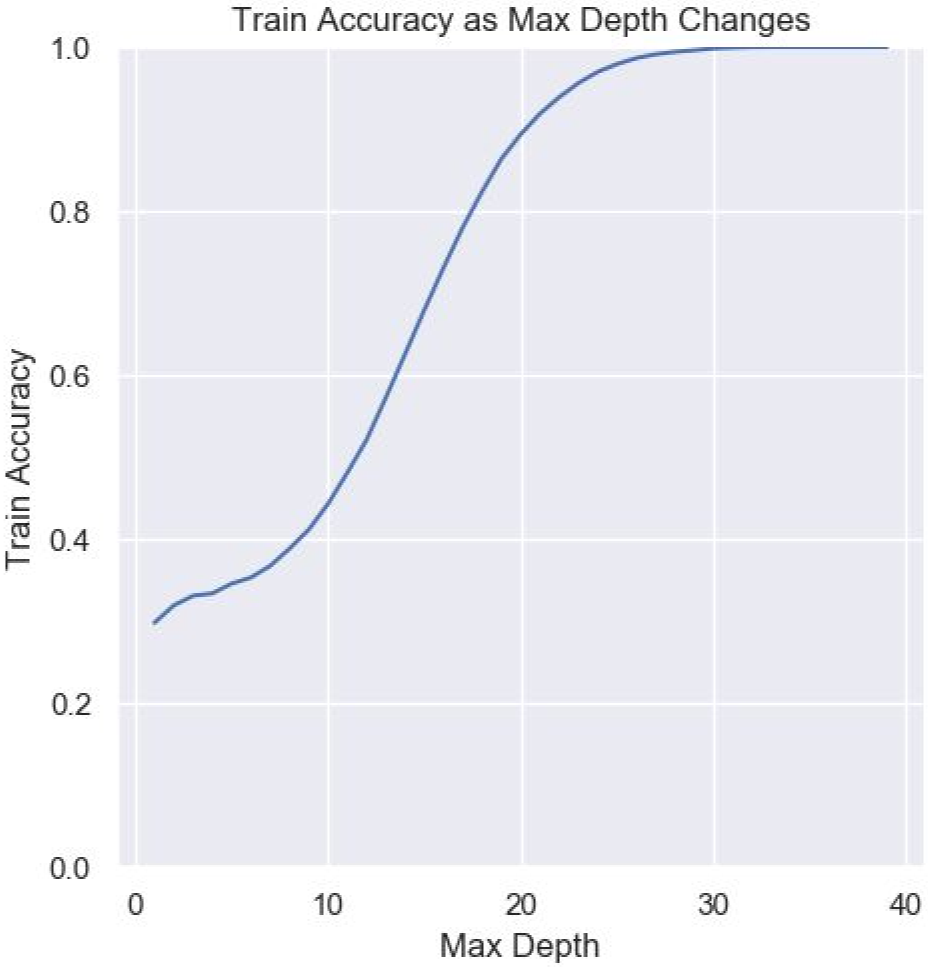


    b. We also explored the relationship between features and the label. We observed that there is no obvious relationship between the features and the label. That is, we can not make linear predictions on the label from any of the features. This is why machine learning comes into place: to help us better predict the response using all other columns we have.

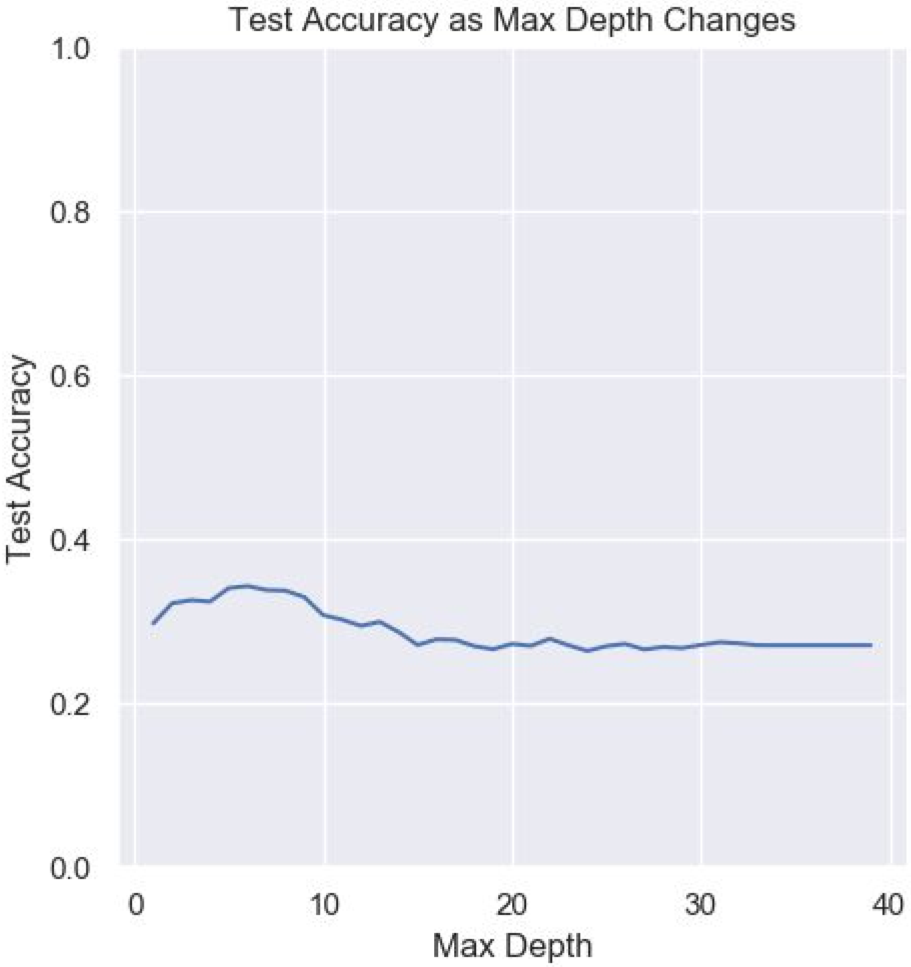


c. We also explored the relationship between test accuracy and training accuracy and one of the hyperparameter "max depth". We did that by changing the hyperparameter "max depth" from 1 to 40 and collecting and plotting the training and testing accuracy along the way.

   i. We can see that the training accuracy increases as the max depth increases. It finally reaches 1 when max depth is around 30. Note that this is so-called overfitting and does not give a good accuracy.



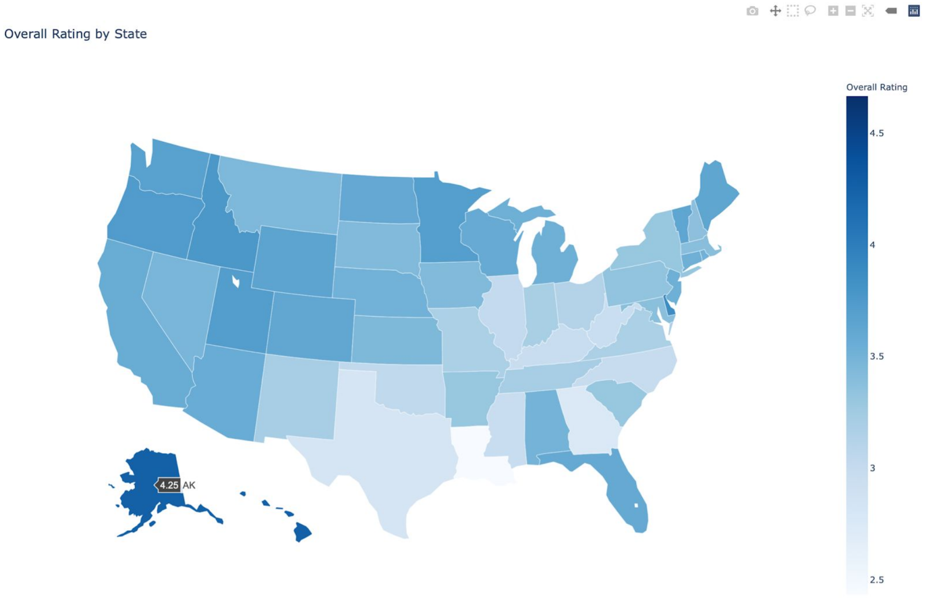
ii. We can also see that the test accuracy first increases with the max depth and then decreases and becomes almost static as the max depth keeps increasing. This is consistent with what we expected because both underfitting and overfitting would not give a nice model. We can observe from the graph that the max depth of 7 or 8 might be a good choice for this hyperparameter because it gives the best test accuracy in our graph.



d. If we guess the level of ratings, the expected accuracy would be ⅕, which is 0.2. We notice that all 40 models we have trained above result in test accuracy that is larger than expected accuracy from guessing. This means that the machine learning model is effective and gives better results than guessing.

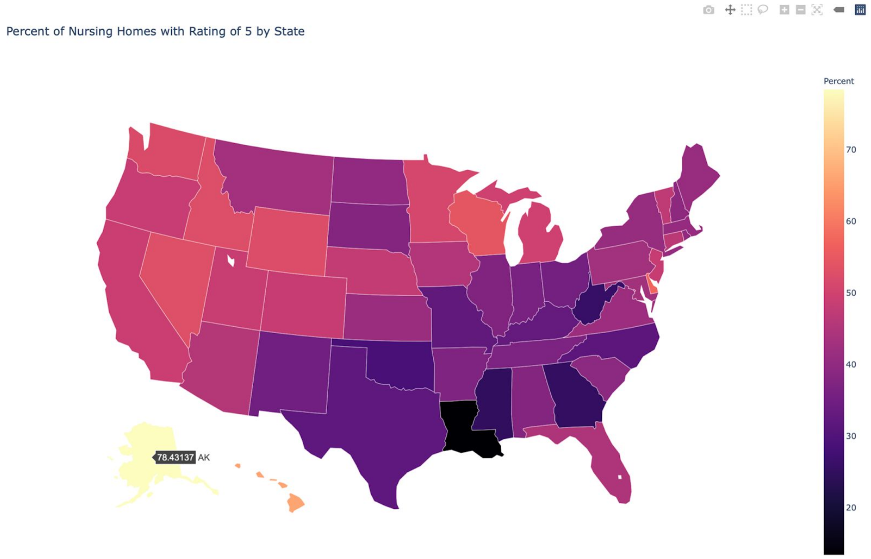
5. Use geospatial data to find the density of high quality and best rated nursing homes in the nation and correlate that with the number of residents in nursing homes in the US. Does a higher demand mean higher rated care?

To answer this question we did some data analysis with the groupby function in pandas and then using plotly, a new package, visualized the result of the analysis by state. As found from our data analysis and visualized with the maps, the state with the highest average overall rating is Alaska, indicating the location in the U.S. with the highest quality of care. Then in general, that state sin the west has a higher quality of quality of care as shown with higher average overall

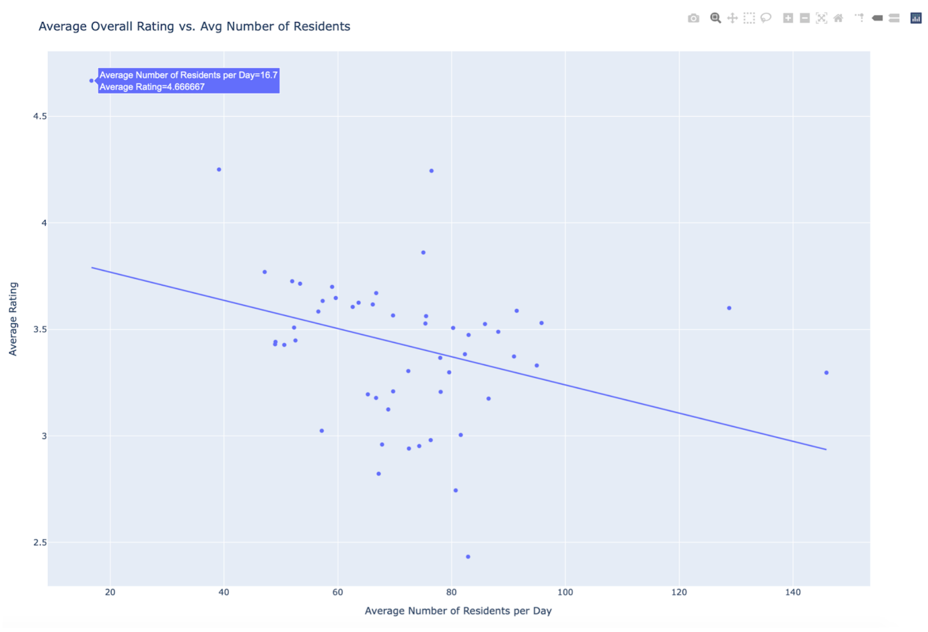


rating.

The state with the highest percentage of nursing homes with rating 5 is also Alaska, supporting the conclusion that Alaska is where the highest rating nursing homes are most dense. And following inference from above, the states in the west side generally have a higher percentage of high quality nursing homes (with a Rating of 5) as indicated by the higher percentages.



Then finding the average of quality of care of each state and compare to the average of amount of resident for each state, there is a general pattern as the average number of resident increases, the overall rating, therefore the quality of care decreases. The answers our question that a higher demand actually correlates with a lower rated care, which is not surprising, since with more residents and people to care for, the caregivers' workload increase and therefore less time and inevitably lower quality care towards each individual.



Challenge Goals:

Our challenge goals are “**machine learning**”, and “**new library**”.

Machine Learning

1. We predicted the overall rating of a nursing home using other features such as the ownership, number of certificated beds, average number of residents per day, etc.
2. We built our model by randomly splitting the overall data set into a train set and a test set(80% vs 20%).
3. Since our predicted variable is a categorical variable, we plan to solve this problem using classification. Specifically, we plan to use decision trees to do the classification.
4. We explored the hyperparameter “max depth” and compared the accuracy between different models with various parameters. We also plotted a graph showing the change of complexity versus the accuracy.
5. Since we are going to predict the overall rating of a nursing home over a scale from 1 to k, we may use multi-class classification. In this case, the test accuracy is at least 1/k, so we are confident that our model is not randomly guessing the classes.
6. We also compared the test scores between different models and see which one is the best based on the score.

New Library

7. We used a new library called Plotly, a very powerful library for plotting that includes functions for various types of charts. We used it to plot the distribution of the response variable, which degree of impacts for each variable and ones with the heaviest influence using a heatmap. We also used geospatial data to plot and find the density of high quality and best rated nursing homes in the nation and correlate that with the number of residents in nursing homes in the US. We also took advantage of the interactive features available in plotly, specifically making a Mapbox

Choropleth Map with plotly, to graph and display the average health care quality (aka. rating) for each state.

Work Plan Evaluation:

We used github to collaborate and one of the group members worked on the machine learning part and the rest worked on the data visualization part.

a. Preparation (7~10 hrs)
    i. Check NAs and clean the data
    ii. Plot the values of overall rating, which is our response variable
    iii. Plot the correlations between features (maybe using seaborn) and may need to remove highly correlated pairs
    iv. One-hot encoding for categorical variables
    v. Split the data into training set, validation set, and test set
b. Machine learning (3~4 hrs)
    i. Use classification to predict the response variable
c. Result analysis (4~5 hrs)
    i. Calculate the accuracy of the model
    ii. Change some hyperparameters such as the depth of the tree and do another train
    iii. After trying some hyperparameters, choose the one that give the best result and explain how features are influencing the response variable
d. Data visualization (5~6 hrs)
    i. Plot the data on maps and see the geospatial distribution of the nursing home and their corresponding rating

Our plan is accurate for the machine learning part but data visualization took more time than expected. This is because we are using a new data set for plotting which requires some time to explore.

Testing

a. For the machine learning part, we did testing by splitting the data into the training set and the testing set. We verified that our models are good because the test accuracies are larger than 0.2 which is the expected accuracy if we guess the result. We also verified the correctness of the code by looking at the trend in the train accuracy.

Collaboration

a. Qiaoxue implemented the machine_learning.py and analyzed the graphs and plots from the part.

Report

- 4-6 pages of text long, but there are no fixed upper or lower bounds on its size. Annotate any visualizations with the method used to produce them.
- Visualizations should add to report's narrative and should be explained in your analysis.
- Plots should be included in report, but should also submit plot images produced by code.
- Label your sections, permitted to write additional sections as well.

1. **Title and author(s).**
2. **Summary of research questions AND RESULTS.** Repeat your research questions in a numbered list.
   After each research question, clearly state the answer you determined. Don't give details or justifications yet — just a brief summary of the answer.
3. **Motivation and background.**
   Same information as Part 1 unless otherwise indicated by feedback
4. **Dataset.**
   Same information as Part 1 unless otherwise indicated by feedback
5. **Methodology (algorithm or analysis).** It is likely that you will come back and refine this section after implementing your project.
6. **Results.** Present and discuss your research results. Treat each of your research questions separately. Focus in particular on the results that are most interesting, surprising, or important. Discuss the consequences or implications.
   Interpret the results: if the answers are unexpected, then see whether you can find an explanation for them, such as an external factor that your analysis did not account for. A good report not only presents the results, but gives an interpretation of them to the reader.
   Include some visualization of your results (a graph, plot, bar chart, etc.). These plots should be created programmatically in the code you submit. If you have to create plots by hand using a program like Excel, you must provide a good reason why it was not possible to create the plot you wanted using Python.
7. **Challenge Goals.** In this section, you should outline which of the challenge goals from Part 0 you think your project completes and why you think so. Be specific in stating which challenge goals your project meets explicitly.
   It is acceptable for you to scale back, or to expand, the scope of your project if necessary. It's better to do a great job on a subset of your original proposal, than to do a bad job on a larger project. If you have to scale back, then explain why the task was more difficult than you estimated when you wrote your proposal. This will help you to make a better estimate for your next project. It will also convince the course staff that you have done an acceptable amount of work for CSE 163. If changes to your project caused you to meet different challenge goals than you originally proposed, that is also okay. However, you should keep in mind that your mentor gave you feedback on your project in the context of your original goals so you should really make sure your changed project meets your new goals.

8. **Work Plan Evaluation**. Include your work plan from Part I (all parts) and evaluate it. Specifically, answer how accurate were your work plan estimates from Part I? Why were your estimates good or bad?
9. **Testing.** You should make some attempt at testing your code to increase your certainty that your analysis is correct. In your report, describe how you tested your code. Did you use asserts? Smaller data files? Be sure to submit your tests and any testing files when you submit your code. You must also have artifacts or evidence from your testing (as in you cannot just say you used print statements and then remove them all from your code when you turn in) Make sure you tell us why we should trust your results!
10. **Collaboration.** State which students or other people (besides the course staff and your group mates, if any) helped you with the assignment, or that no one did. Did you use online resources to help you? If so, what were they?