Edwin Ng

Nursing Home Staffing Hour Analysis Final Report:

Problem Statement:

Abuse of the elderly, specifically in institutions such as nursing homes or long-term care facilities, is a serious problem that is sadly observed around the world. According to a “2017 review of 52 studies in 28 countries” it is “estimated that over the past year 1 in 6 people (15.7%) aged 60 years and older were subjected to some form of abuse (1)” (WHO). Furthermore, studies indicate that 2 in 3 staff members of institutions such as nursing homes or long-term care facilities have admitted to perpetrating some form of abuse of elderly residents within the last year (WHO).

On a more personal note, a close friend of mine is also a certified nursing assistant (CNA) at a well known retirement home in my state. While he finds the occupation rewarding, it is also clear that working in a retirement home as a caregiver is also a mentally draining and challenging job to say the least. Faced with residents that mistreat or even harass/assault the staff (either as a result of their medical conditions or not) and problems such as understaffing and underpaying - it is no wonder that “nursing homes, experienced turnover of 52% of nursing staff each year.” (The Consumer Voice, 2022)

That being said, I want to explore the relationship between nursing home staff and resident care. More specifically, I want to explore the relationship between nursing home staffing rates and nursing home care ratings. This could also lead to more investigations of the relationships between staffing rates and other rating factors such as incident reports, fines, penalties, etc.

How can we use nursing home staffing rates to uncover its impact on nursing home quality measures such as overall care ratings, health inspection ratings, incident reports, and fines/penalties such that we can make staffing recommendations to local nursing homes within the next year to reduce resident mistreatment and improve staff QOL?

Edwin Ng

Data Wrangling:

The nursing home quality staffing dataset imported from Kaggle initially contained 15640 entries and 82 columns. While it contained lots of useful data, it also contained a lot of data that did not pertain to our question. Since our question mainly involved staffing hours, I went ahead and subsetted our dataset to only contain rows with information about staffing hours and our responding variables.

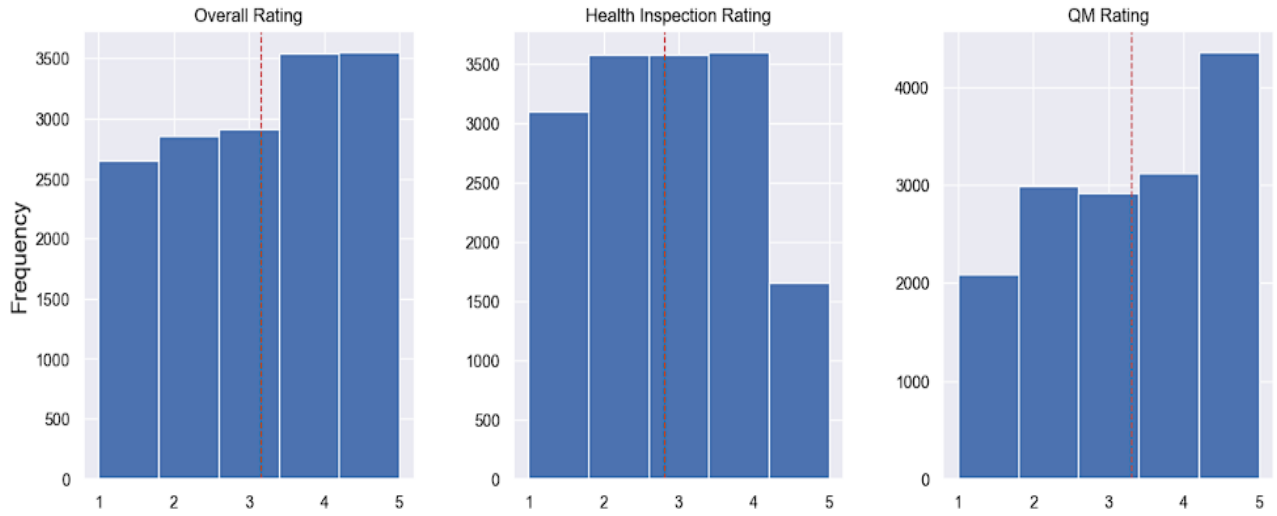
After this, the next step to this phase was to ensure that all our columns contained the right kind of information. From a quick glance it was clear that not all columns contained the right types of data. For example, the column “total amount of fines in dollars” contained the object data type instead of a numeric data type because the entries contained a dollar sign ($) symbol and had to be removed before it could be converted into the numeric data type of float64. After ensuring column data type integrity we made sure that there were no duplicate entries within the dataset before moving onto the issue of null values.

When it came to the null values within our dataset, we had to first eliminate all the rows where our responding variables - Overall Rating, Health Inspection Rating, QM Rating - were null because we are unable to impute these values since it would skew our results. This is especially due to the fact that there are (and were) so many other variables present within our dataset. After removing said rows, the only remaining null values were 2 rows where the 'Reported Physical Therapist Staffing Hours per Resident Per Day' value was missing. This was dealt with by replacing the null value with the mean value for this metric across the dataset. In the end, our dataset’s final shape after wrangling was 15183 entries across 41 columns.

Exploratory Data Analysis:

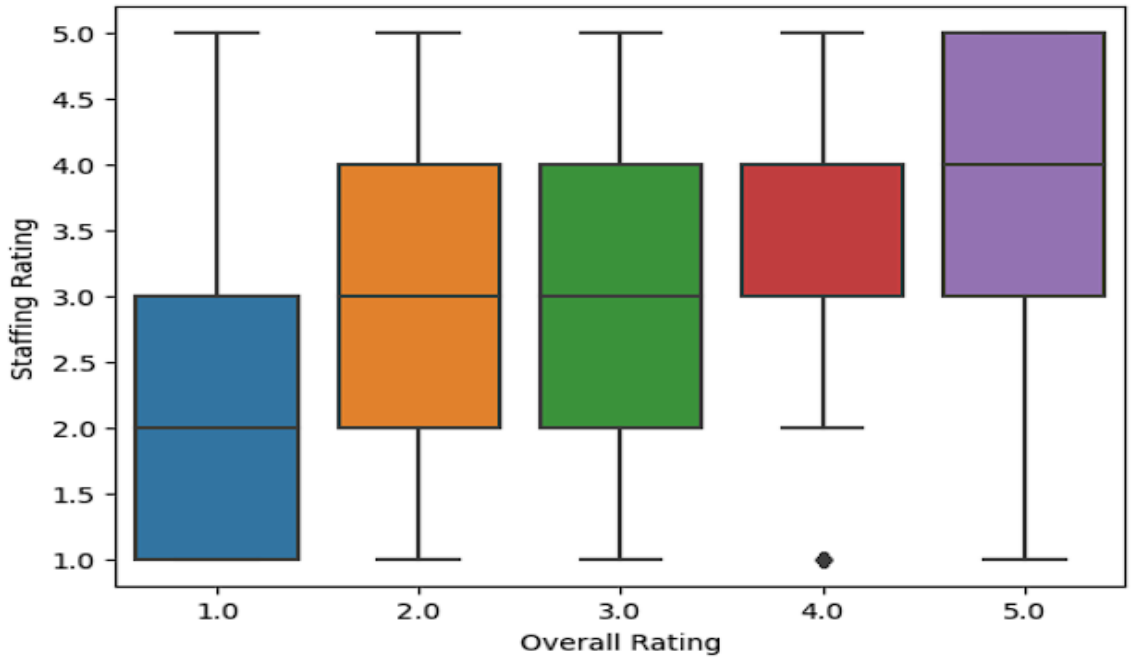
We first took a look at the three different metrics that indicate the quality of a nursing home according to the dataset - overall rating, health inspection rating, and QM rating. We explore these 3 metrics by getting a sense of their distribution using a histogram.

Edwin Ng



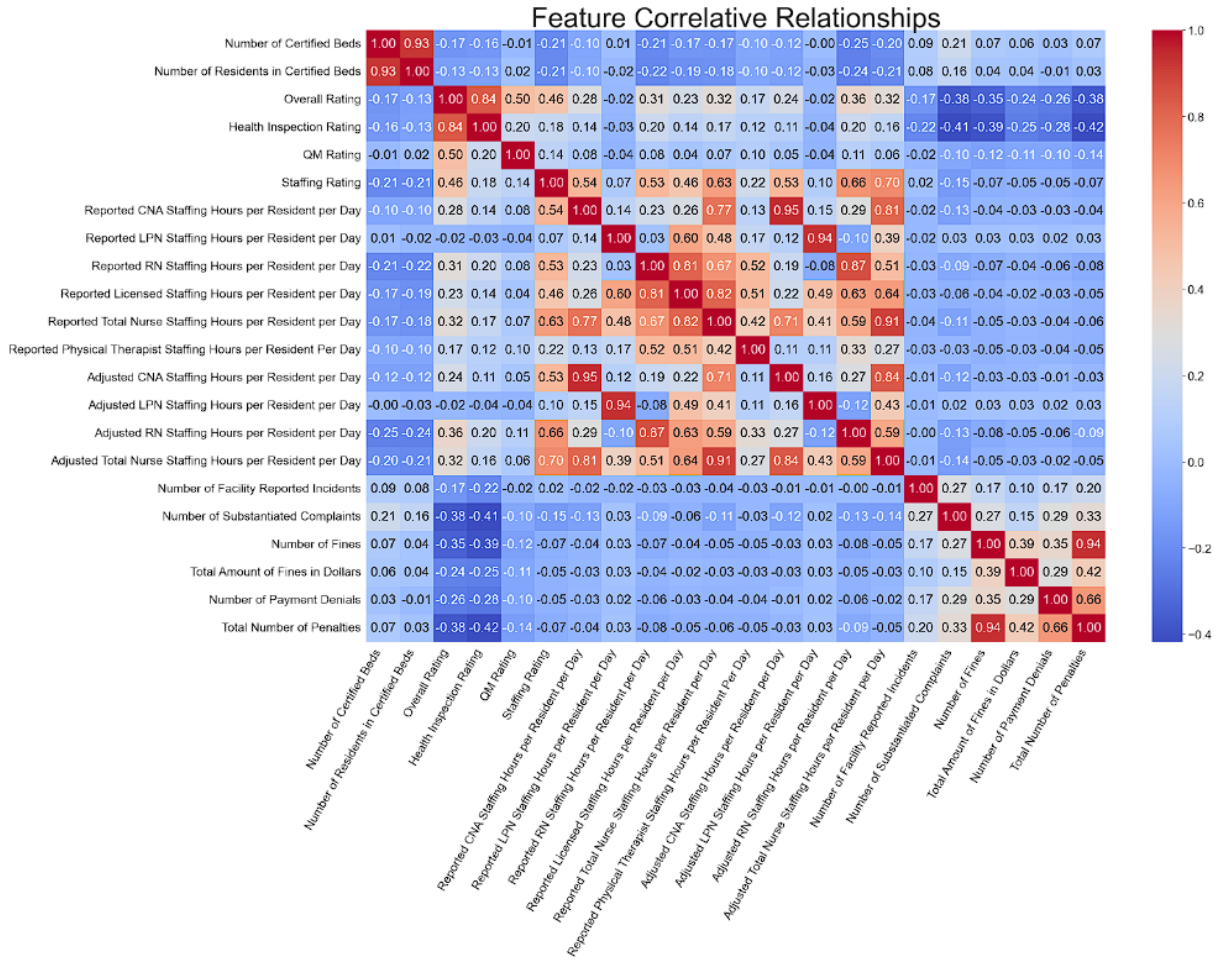
Overall rating, as the name might suggest, takes into account all other factors and metrics and produces an overall score for the facility. Health inspection rating is the score that the nursing home earns for each of the health inspections throughout the year. Finally, QM rating - standing for quality management - is a combination score assessed through the measure of quality of care provided at the facilities as well as complaints received.

On top of this, we take a broad level look at how staffing rating - a measure of how overall well staffed a nursing home is - affects overall rating.



Though not entirely super informative, we can already see that as staffing rating increases, there is a positive trend in our overall rating as well, supporting the idea that better staffing leads to better overall ratings.

While there are many relationships we could graph, it would eventually become an aimless effort that would take much more time than needed. As such, we constructed a heatmap to point out notable relationships with our desired responding variable and graph them out.



Notable variable relationships:

Positive Relationships

- Health inspection rating / Overall Rating: *0.84*
- Adjusted Total Nurse Staffing Hours per Resident per Day / Staffing Rating: *0.7*
- Adjusted RN Staffing Hours per Resident per Day / Staffing Rating: *0.66*
- Reported Total Nurse Staffing Hours per Resident per Day / Staffing Rating: *0.63*
- Reported CNA Staffing Hours per Resident per Day / Staffing Rating: *0.54*

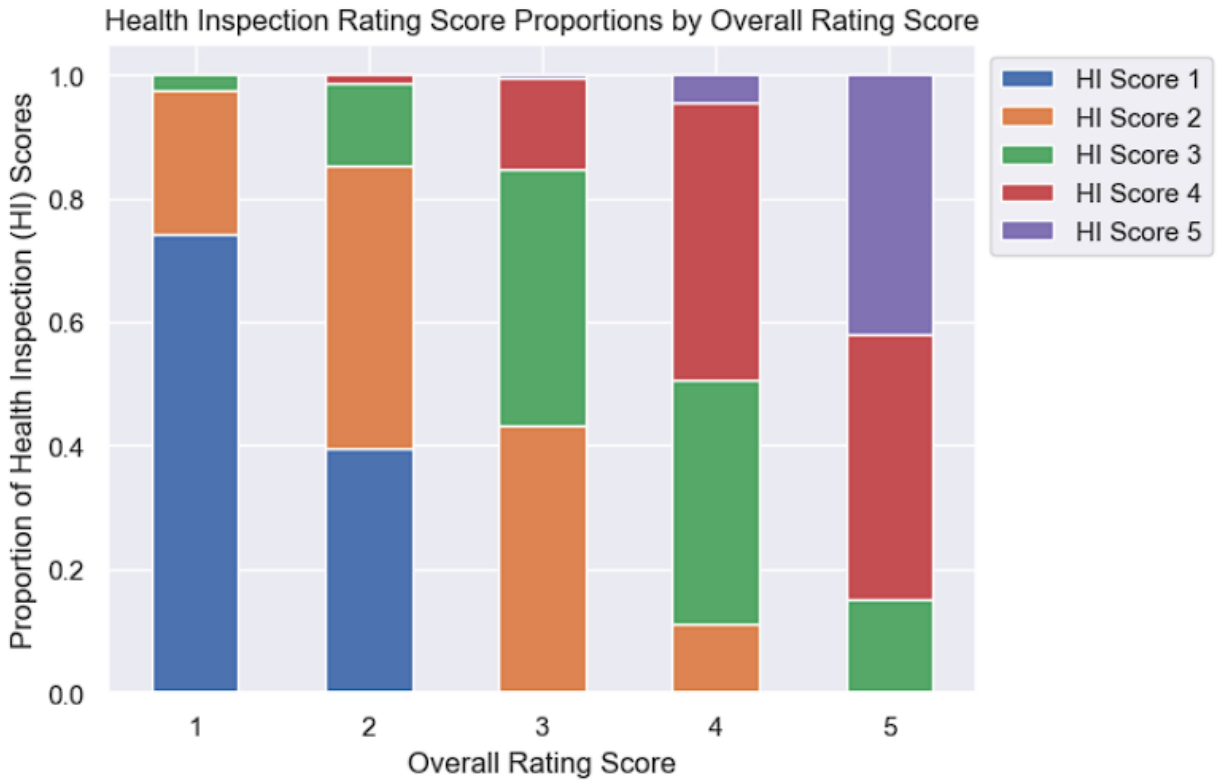
- Adjusted CNA Staffing Hours per Resident per Day / Staffing Rating: *0.53*
- Reported RN Staffing Hours per Resident per Day / Staffing Rating: *0.53*
- Overall Rating / QM Rating: *0.5*
- Overall Rating / Staffing Rating: *0.46*

Negative Relationships

- Health Inspection Rating / Total Number of Penalties: *-0.42*
- Health Inspection Rating / Number of Substantiated Complaints: *-0.41*
- Health Inspection Rating / Number of Fines: *-0.39*
- Overall Rating / Total Number of Penalties: *-0.38*
- Overall Rating / Number of Substantiated Complaints: *-0.38*
- Overall Rating / Number of Fines: *-0.35*

Some of the more notable relationships that we graphed included:

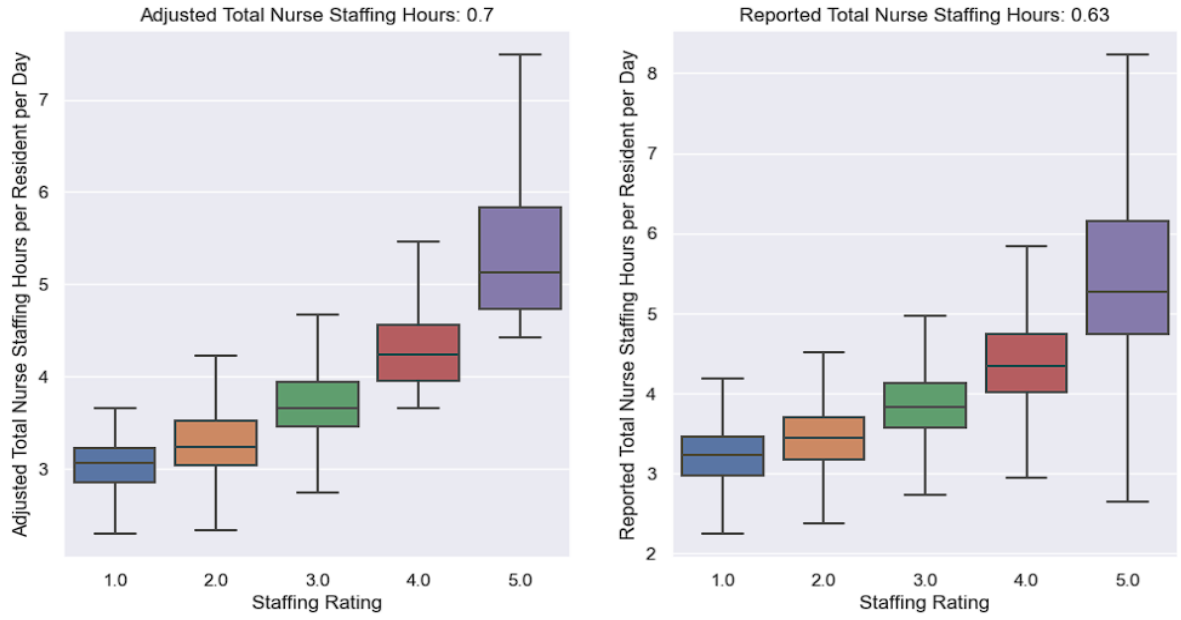
Health inspection rating score proportions by overall score.



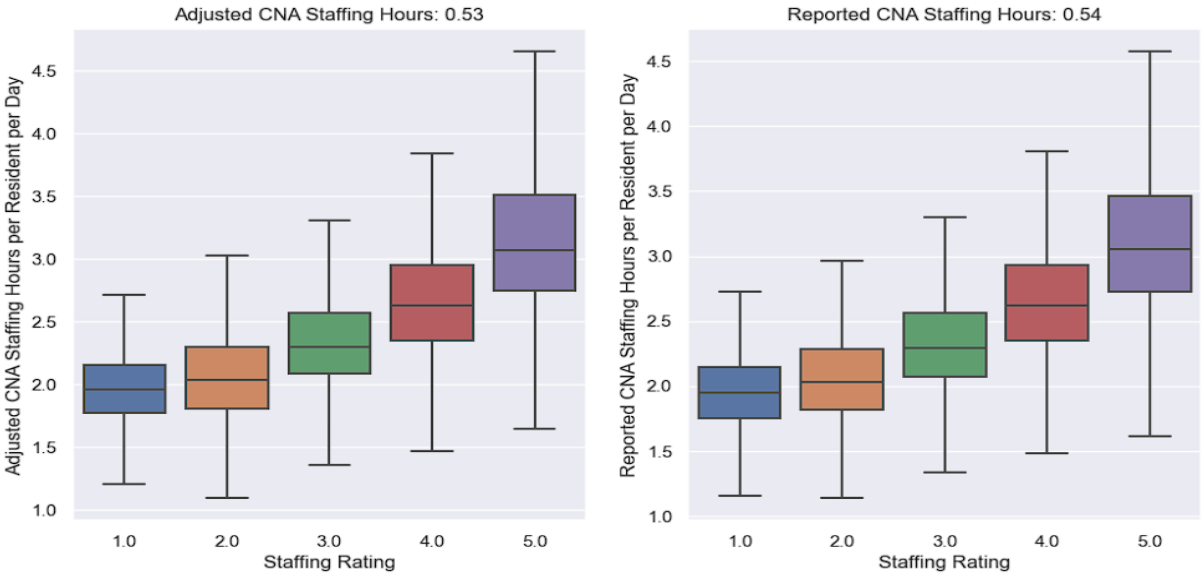
Here, we visualize the health inspection score proportions of each overall rating score category. For example, we can see that a little over 75% of nursing homes that scored an overall rating of 1 had a health inspection score 1 as well. This visualization highlights the importance of maintaining a decently high health inspection score if the

nursing home wants to score highly on their overall rating. In fact, from this graph we can even see that there is not a single nursing home with an overall rating of 5 that had a health inspection score of 2 or lower - and even then less than 20% of nursing homes in that category scored a 3 on their health inspection as well.

Adjusted and Reported Nurse Staffing Hours vs. Staffing Rating

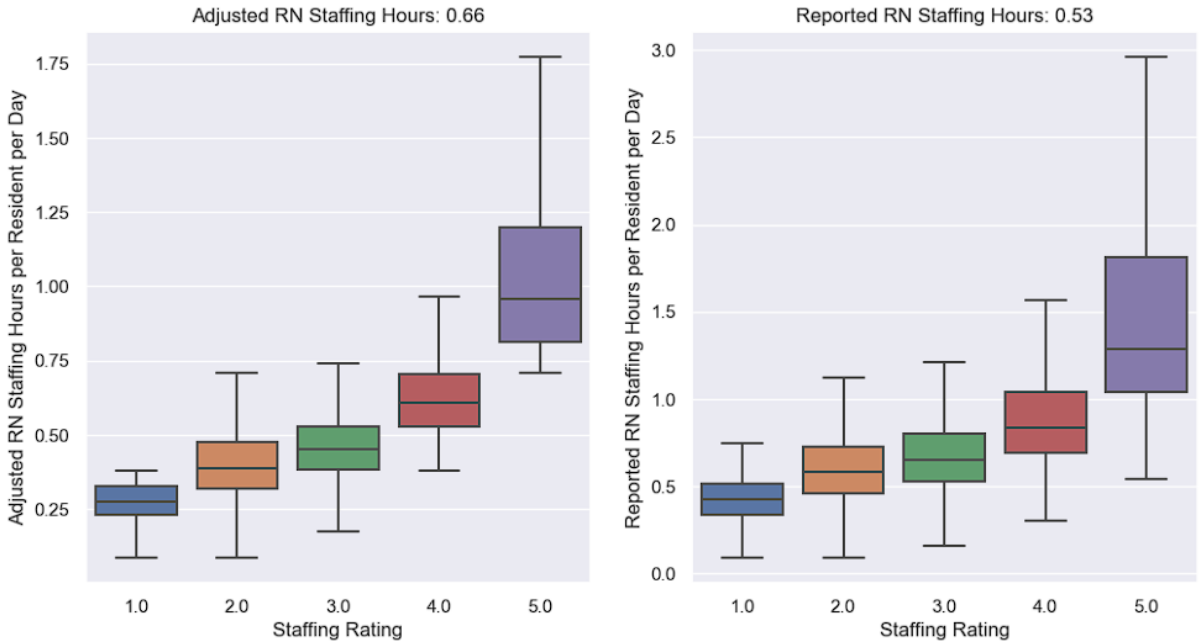


Adjusted and Reported CNA Staffing Hours vs. Staffing Rating

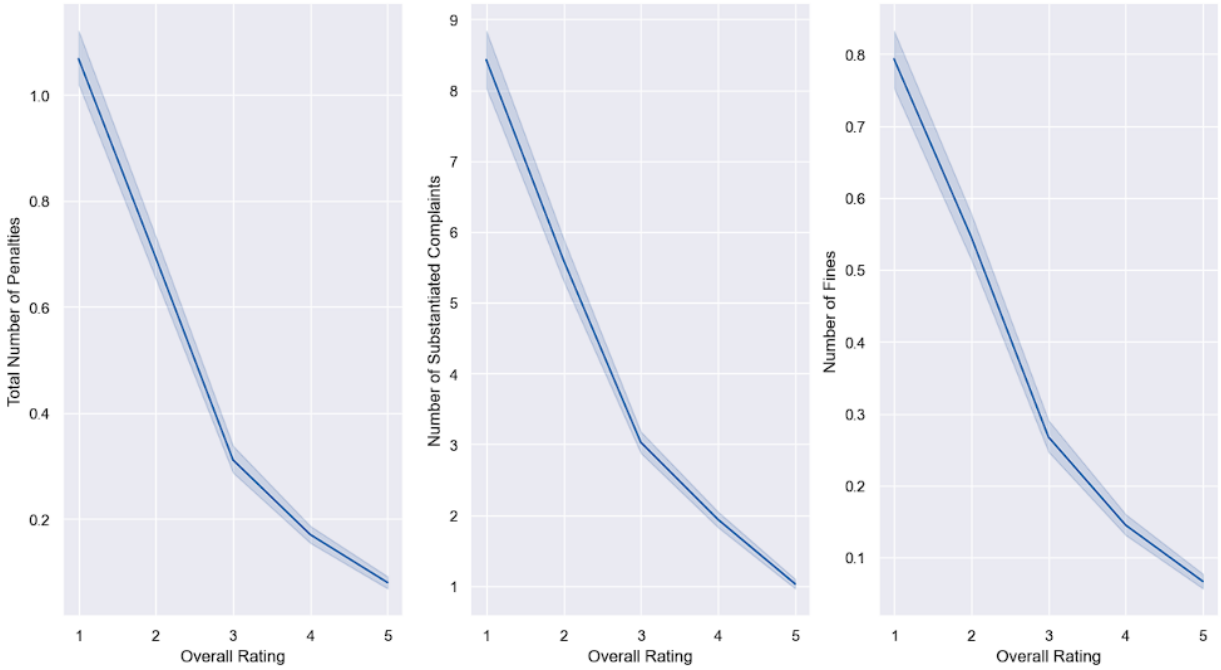


Edwin Ng

Adjusted and Reported RN Staffing Hours vs. Staffing Rating



Overall Rating vs. Total Number of Penalties / Substantiated Complaints / Fines:



Edwin Ng

Exploratory Data Analysis Conclusions:

From what we could see from our exploratory data analysis, overall rating was strongly positively correlated with health inspection score and moderately positively correlated with the other two overarching quality metrics, QM rating and staffing rating. QM rating being a score given to the nursing homes based on quality measure scores. Staffing rating is a score given to the nursing home based on the staffing levels of the facility.

Since our original project proposal was focused on the staffing aspect of the nursing homes, that was our immediate focus. Upon investigating the staffing rating score, 3 different staff metrics stood out in terms of their correlative relationship with the staffing rating. These three were total nurse staffing hours, followed by specifically nurse and certified nursing assistant (CNA) staffing hours. As expected, the total nurse staffing hours had the strongest positive correlation with the staffing rating, however it was clear that registered nurses and CNAs had the strongest effect on staffing rating scores when compared to other roles such as LPNs and physical therapists.

As for the health inspection metric, I noticed that aside from the obvious strong positive correlation with overall rating, the health inspection metric had moderate negative correlative relationships with the number of penalties, substantiated complaints, and fines. As expected these relationships carried over to overall rating as well.

Modeling:

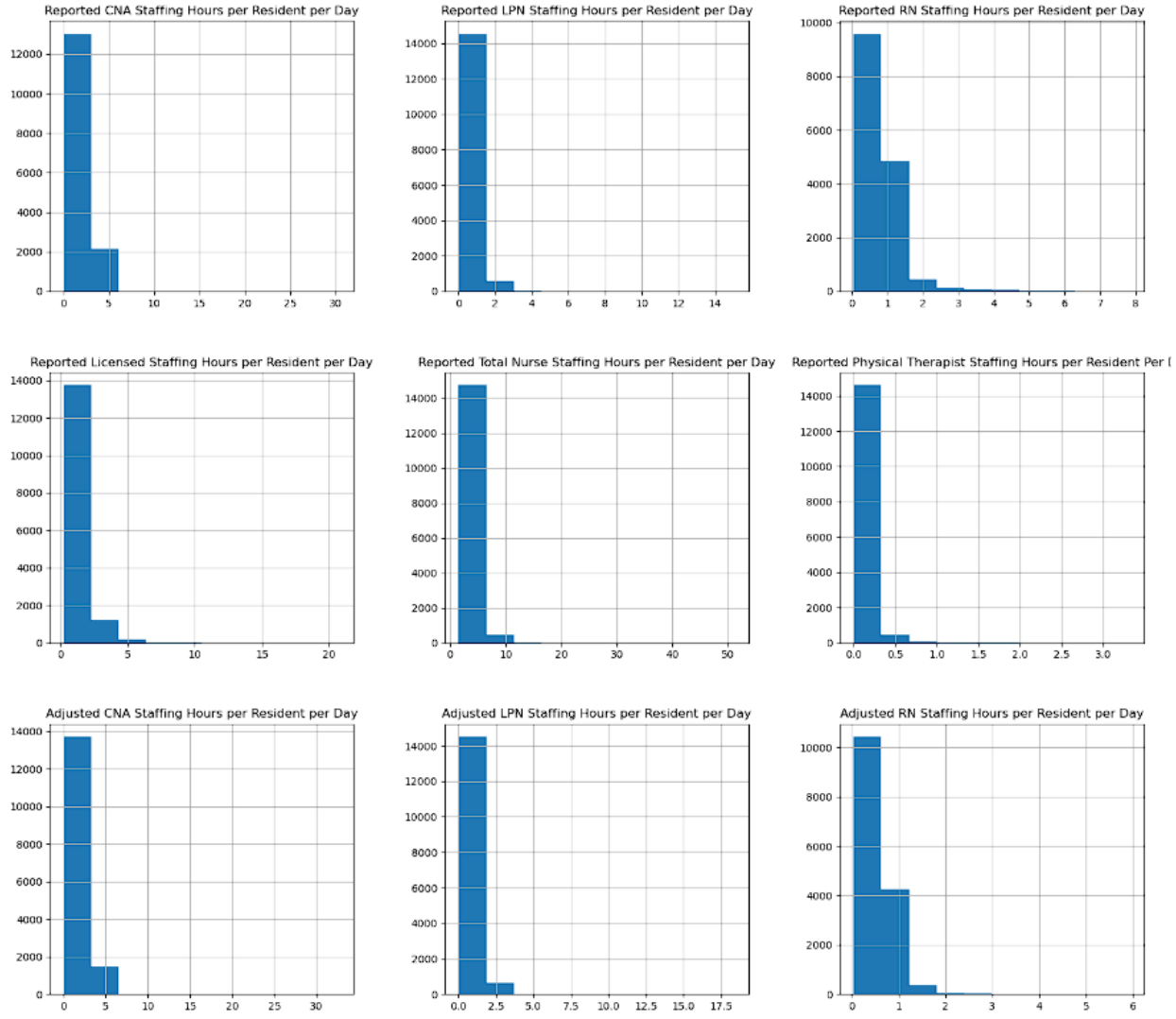
Creating Dummy Variables:

We first began the modeling phase of this project by taking our cleaned and preprocessed data and turning the categorical variables into dummy variable columns. When it came to choosing important categorical variables to work with, two variables came to mind with this dataset: Ownership Type, and Provider Type. After this step was carried out, our dataset went from 27 columns to 41 columns. We then removed the federal provider number and provider name columns since they would not be helpful in our analysis and also had non-numeric entries - which would be problematic when working with our models.

Edwin Ng

Standardizing Values:

Since we would be working with feature importances and the K-nearest neighbors model later on we then moved on to standardizing the values within our dataset.



From a quick glance of our distributions, even though there are some outliers that sat far outside the mean, it was still helpful to normalize our data. Since there weren't a lot of outliers, I chose to normalize the data instead of transforming it on a logarithmic scale. This centered our data around the mean and measured data points by their standard deviations about the mean instead.

Splitting the data:

From here, we split our data on an 80/20 train-test split - focusing only on overall rating since that was the focus of our project. Moving onto our model selection, we chose the following three models:

Linear Regression:

- The linear regression mean squared error: **0.2126697509640798**
- The linear regression model R squared: **0.8948769714186898**

K-Nearest Neighbors:

After using GridSearchCV to estimate our best parameters, we found that the best number of neighbors was 172 - which yielded the results:

- The K Nearest Neighbors model mean squared error: **2.7925584458347052**
- The K Nearest Neighbors R squared value: **0.30260125123477116**

Random Forest:

Again using GridSearchCV and the following parameter grid:

```
params = {
   'bootstrap': [True],
   'min_samples_leaf': [1, 2, 3, 4, 5],
   'min_samples_split': [2, 4, 6, 8, 10],
   'n_estimators': [100, 200, 300, 1000]
}
```

We found that the best parameters for our model were:

RandomForestRegressor(min_samples_leaf=2, min_samples_split=8)

Earning a score of:

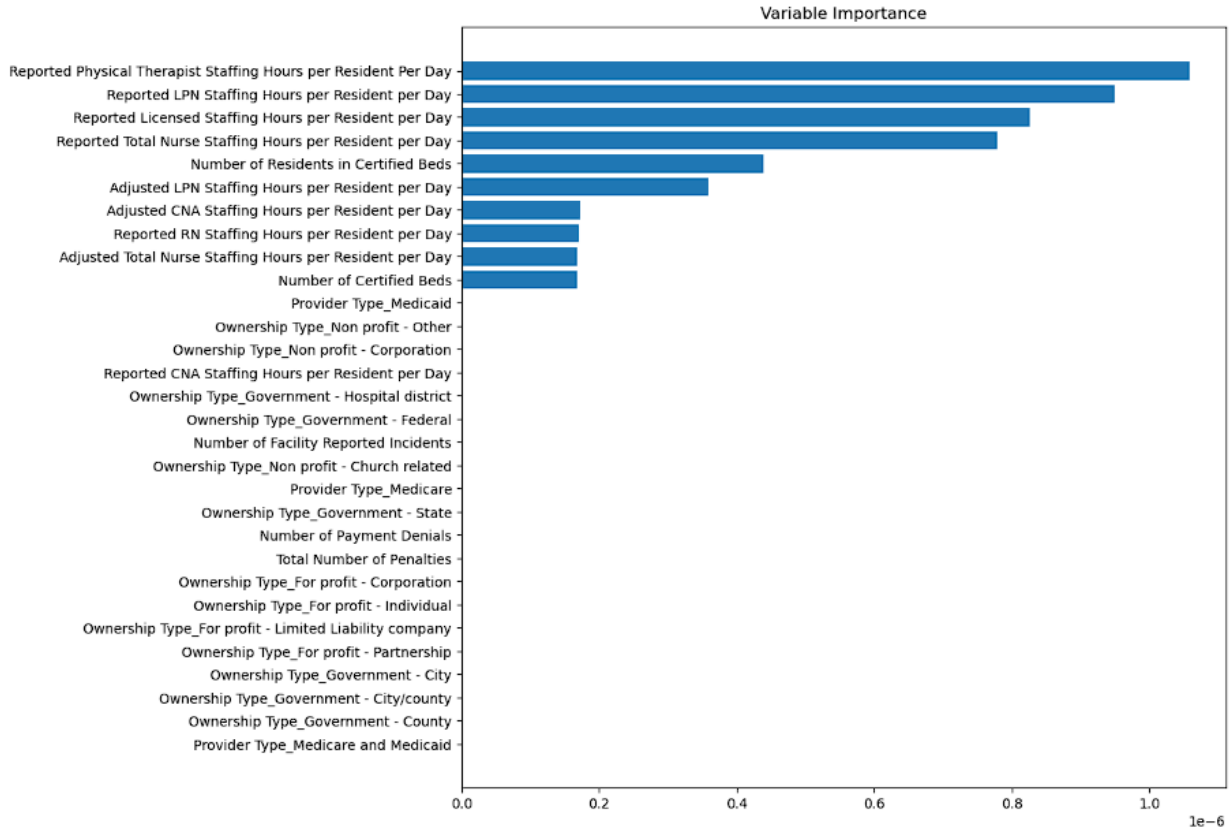
- The K Nearest Neighbors model mean squared error: **0.0013419260176346691**
- The K Nearest Neighbors R squared value: **0.999336683630529**

Modeling Conclusions:

Among the three models that we tried, it is clear that the Random Forest model performed the best. Surprisingly the Linear regression model performed better than the K Nearest Neighbor model. This might be because the shape of the data was more linear than we initially assumed and better fit the linear model rather than the K Nearest

Neighbors model. Finally, the Random Forest model performed the best, most likely due to its ability to accommodate multiple types of data within the datasets. On top of this, the model is able to modify many different parameters to best fit the dataset. This leads to the model performing the best out of the set of three that we tried.

As for our feature importances, this is what we found:Mo



Ranked from most important to least, our top 10 most important features were:

1. Reported Physical Therapist Staffing Hours per Resident per Day
2. Reported LPN Staffing Hours per Resident per Day
3. Reported Licensed Staffing Hours per Resident per Day
4. Reported Total Nurse Staffing Hours per Resident per Day
5. Number of Residents in Certified Beds
6. Adjusted LPN Staffing Hours per Resident per Day
7. Adjusted CNA Staffing Hours per Resident per Day
8. Reported RN Staffing Hours per Resident per Day
9. Adjusted Total Nurse Staffing Hours per Resident per Day
10. Number of Certified Beds

At first this kind of surprised me because the first two categories were the same as those immediately flagged by our exploratory data analysis. They were also not what

immediately come to mind when you ask someone to think of the most important people in a nursing home either. However, with a little more thought, it becomes clear from these feature importances - that comfort is the prevailing factor when it comes to overall rating. After all, physical therapists help with ongoing treatments of injuries and pre existing physical conditions while licensed practical nurses (LPNs) are responsible for basic patient care and comfort. In fact, LPNs are so important that both their reported and adjusted staffing hours came up in the top 10. Otherwise, the rest of the features listed are in line with our expectations when it comes to staffing. However, some features that surprised me were the number of residents in certified beds and the number of certified beds. While they were surprising because these columns did not peak any interest in earlier stages of this project, it makes sense that they are so important because they determine the amount of per-patient care possible in each nursing home. Too many residents and the amount of care gets diluted amongst the many filled beds.

Further Investigations:

While my project so far has enabled me to get a deeper understanding of what goes into the care and operations behind nursing homes, and what makes a good nursing home good, it has also left me with more questions. For example, while it is obvious that the more staff you have on hand to care for your patients the better, there has to be a limit. Therefore what is the ideal ratio of patients to staff such that staff are not overworked and care quality is optimal?

It also seems from our feature importance list that not all roles within a nursing home might are created equal when it comes to patient care and nursing home quality. With this in mind, what is the ideal blend of staff? What ratio of each profession within a nursing home leads to the best quality? Given a realistic staffing budget how can we create the best team? It seems we have just scratched the tip of the iceberg with this little project!