Name: Jolene Branch  
Date: 11/12/2019  
Title: Predicting Quality of Nursing Homes   
  
**Section 1 – Week 9**

* Introduction
  + I am interested in determining if there are characteristics that high quality nursing homes share. There are all kinds of quality metrics available, and I wonder if certain characteristics (or sets of characteristics) can be found that have positive correlation with quality outcome measures. Operators of nursing homes, as well as potential consumers, would be interested in this as a possible clue to infrastructure changes needed at the facility (versus constantly creating action plans to ‘fix’ problems at their facilities) in order to set themselves up to be successful in their continuous quest for quality.
* Research questions
  + What should be the outcome measure? Count of deficiencies? Some sort of adjusted rate?
  + Do organizations that own numerous nursing homes have better quality outcomes than independently owned nursing homes?
  + Does RN staffing have any relationship to quality outcome measures? (because it does in the acute care setting). Such as falls, pressure ulcers, and other health deficiencies
  + Have nursing homes with higher quality ratings set different RN staffing targets?
  + Does owner status, such as ‘not for profit,’ ‘for profit,’ and ‘government owned’ have any relationship to quality outcomes?
* Approach
  + I will use logistic regression to look for explanatory variables that are the best predictors of various clinical quality outcome measures. Having a background in healthcare (and nursing), I will start with characteristics that I believe to have direct effect on quality outcomes.
* How your approach addresses (fully or partially) the problem.
  + My approach would give owners and the government ideas of where to start in improving the quality of nursing home care, but does not control certain aspects, such as RN staffing levels, rurality of location, and condition of the environment within the facility. If say, individual ownership is highly predictive of positive clinical quality measures, I might take that into consideration when picking a nursing home for family, or for when to determining which nursing homes to close within the region (due to, say, lack of funding to keep all the beds open).
* Data
  + I will use the Nursing Home Compare datasets from [data.medicare.gov](https://data.medicare.gov/data/nursing-home-compare). These are the official datasets used on the Medicare.gov Nursing Home Compare Website provided by the Centers for Medicare & Medicaid Services. These data allow you to compare the quality of care at every Medicare and Medicaid-certified nursing home in the country.
    - The [Provider Info](https://data.medicare.gov/Nursing-Home-Compare/Provider-Info/4pq5-n9py) dataset is available in .csv format and contains general information on currently active nursing homes, including number of certified beds, quality measure scores, staffing, and other information used in the Five-Star Rating System. Data presented as one row per nursing home. The dataset contains 86 variables, some of which are derived values.
    - The [Health Deficiencies](https://data.medicare.gov/Nursing-Home-Compare/Health-Deficiencies/r5ix-sfxw) dataset is a .csv downloadable file with 19 variables, including the nursing home that received the deficiency, the inspection date, scope and severity, current status of the deficiency, and the correction date. Data are presented as one deficiency per row.
    - The [Star Ratings](https://data.medicare.gov/Nursing-Home-Compare/Health-Deficiencies/r5ix-sfxw) are a list of a variety of averages for each state or territory as well as the national average, including each quality measure, staffing, find amount and number of deficiencies. Each row displays a specific state or territory, the associated measure, and average.
* Required Packages
  + I will need to use ggplot2, Rtools, dplyr, pastecs, Mmisc, and QuantPsyc.
* Plots and Table Needs
  + I might find some initial direction by creating simple scatterplots off the Provider Info dataset. I see the website that is the source for the data has a way to generate charts
* Questions for future steps.
  + I am going to have to learn how to combine the variables from multiple datasets

**Section 2 – Week 10**

* How to import and clean my data
  + I downloaded the Provider Info dataset from nursinghomecompare.gov via Kaggle. Column names are not intuitive. There's a metadata file with labels for each of the variable names. They are in separate columns. There are too many columns to manually change them.
  + I transposed my current column names, then used the metadata sheet (from a separate download from the same web site) to do a VLOOKUP to get the actual variable names. This worked for all but five that didn't have a corresponding value in the VLOOKUP table. For those I used their original, non-intuitive column name.
  + Each time I saved the changes in the .csv file, I got a warning that the changes in Excel might not all be compatible with the .csv file format. That got me worried about making too many changes to the dataset before I imported it into R, so I moved all the other ‘working’ tabs to a new Excel workbook and just saved the actual dataset (with improved column names). That is what I imported into R using read.csv. Worked fine.
  + I started following the steps as demonstrated in the DataCamp ‘Cleaning Data in R’ module. I used class(), dim(), names(), and str(). This data set is over 15,000 rows long. I think I could narrow my scope to just Wisconsin nursing homes and have my work actually be more valid – because the healthcare environment varies across regions. Even within the state, some markets are rural and some are urban. Heck, maybe I should do it by telephone area code. (Entire state of Wisconsin has only three area codes).
  + I called head() and tail() on the dataset, and was surprised to see the words in each column name all separated by periods.
  + I should replace all the periods in the column names with spaces, but was not sure if I needed to do it now or not.
  + I called glimpse() and saw a much easier to view and understand version of str().
  + I called summary() to see a summary of the distribution of each column. The variables that I was going to consider to be response variables all seem to have super low medians and means. Like 0.9s and 1.0s. (These are things like count of fines, count of facility-reported incidents, and count of substantiated complaints).
  + The data appears to be tidy, and I don’t see duplicate observations. I confirmed this using conditional formatting in Excel (because I thought it would be quicker). I noticed that the Federal Provider Number seems to be the primary key, and that scattered throughout the dataset are group of a handful of nursing homes that do not appear to follow the Federal Provider Number naming convention (of a five digit number). I also noticed more than one references to “Children” among the provider names. Unless there is a new trend in mixed use facilities (for day cares and nursing homes), I believe I need to eliminate any observations with the words child or children in the Provider Name variable from the data set.
* What does the final data set look like?
  + A table format , 80 columns wide and 15,000+ rows long, displayed in ascending order by the Federal Provider Number. 5652 N/As. Most, if not all the staffing variables have N/As, but they all seem to have the same sum of N/As, (probably because the same 394 facilities did not enter staffing numbers for any of their levels of nursing staff), so maybe no need to start thinking about omitting any data unless I simply omit the observations of the nursing homes with any N/As.
* Questions for future steps.
  + How do I remove all observations that have the word ‘child’ or ‘children’ from the data set?
  + Do I need to keep that information somewhere?
  + Do I need to remove the N/A’s from the staffing variables? (It looks like they all have the exact same number of N/As, so probably not).

**Section 3 – Week 11**

* What information is not self-evident?
  + It is not evident if there is strong correlation between the ownership type and the staffing levels, and the staffing levels and quality rating (or number of incidents)
* What are different ways you could look at this data?
  + I could subset the nine ownership types into just three; ‘for profit,’ ‘non-profit,’ and ‘government.’
  + I could do a facet grid of all variables against the dependent variable, QM Rating.
* How do you plan to slice and dice the data?
  + I don’t want to slice and dice the data. I want to look for the strongest correlations between the independent and dependent variables.
* How could you summarize your data to answer key questions?
  + I could create a table of Pearson correlations between select variables.
* What types of plots and tables will help you to illustrate the findings to your questions?
  + A facet plot of ownership types to QM Rating would be helpful
  + A facet plot of Adjusted Total Nurse Staffing Hours per Resident per Day to QM Rating
* Do you plan on incorporating any machine learning techniques to answer your research questions? Explain.
  + It would be interested to create a model ….the danger in creating a model that shows say, ideal staffing levels (or lowest possible staffing levels) that don’t negatively influence QM Rating or Number of Facility Reported Incidents, is that staffing levels are strong influencers of other clinically significant outcomes, such as pressure ulcer formation and post-operative complications. But once you start trying to control for too many variables, the effect of one on the ultimate outcome measure will become less pronounced.
* Questions for future steps.
  + Are these findings generalizable for the data set containing all 50 states?
  + In cities with both profit/non-profit AND government ownership, are the quality ratings of the profit/non-profit facilities better than in cities that have no government owned nursing homes? (This gets at the inherent concentration of Medicaid residents within government owned facilities and the general public perception that nursing homes with fewer Medicaid residents are able to pay staff more or pay for more staff and are therefore safer and more desirable nursing homes in which to reside).
  + Can my computer even handle that kind of number crunching?

**Section 4 – Week 12 -** A story / narrative that emerged from your data.

* Introduction. Having worked as an RN in the inpatient acute care side of health care for over ten years (which is another way of saying ‘hospital units’), I was interested to see if the same correlations between RN staffing levels and clinical quality exists on the nursing home side of health care. I was aware of the presence of nursing home compare data through my current job, but was not aware that I could access the raw data tables until I started exploring various data sets for this final project.
* The problem statement I addressed.  I was attempting to answer the question “Are there characteristics that higher performing nursing homes share?” This information could be helpful for consumers in selection of nursing homes for self or loved ones, and could potentially lead to a sort of predictor model for how to construct and manage a nursing home with high quality results.
* How I addressed this problem statement. What I really wanted to find was not so much ‘shared characteristics,’ but more like directional correlation. I addressed the problem statement by cleaning up a national data set from data.medicare.gov, sub-setting to retain only nursing homes in the state of Wisconsin, reviewing for outliers, and removing the handful of facilities that had not submitted staffing data. The decision to include only facilities in the state of Wisconsin was based on convenience (to decrease the 15,000+ row data set’s size) and acknowledgement that regional differences in provision of health care, especially related to rural/urban status, could be so vast as to negate any differences potentially observable on a smaller scale. Think regression to the mean.
* Analysis. The QM Rating scale of 1-5 could possibly have hindered attempts to determine correlation. Fortunately for scaling purposes, no facility reported more than 20 incidents in the time period. The slope of the regression line for Number of Facility Reported Incidents and QM Rating looked promising, and prompted further investigation into the correlation of these variables. Using the Pearson correlation coefficient, I found basically no correlation between the number of nursing hours spent on patient care per day and the quality rating of nursing homes. (R2 = 0.4%). The regression line for Nurse Staffing and Number of Facility Reported Incidents had an R2 of 0.02, which meant that nursing hours of care per day can account for 2% of variation in number of facility reported incidents. But even though the p value is quite low (at 0.00568), this predictor model isn’t much better than simply using the mean value of ‘Number of Facility Reported Incidents’. Bringing understanding of the differences in psychological safety among facilities, and knowing the effect that those differences have on staff reporting of incidents, it is altogether possible that the regression line in question has very little basis in reality. (The number of reported incidents is just that; the number of incidents that actually got reported. This number may have minimal correlation to actual incidents).
* Implications.  It appears that the tried and true hours of nursing per patient per day metric that is a predictor of better quality outcomes (and fewer complications) in the hospital side of health care is not an automatic predictor of similar improved outcomes in the nursing home side of things.   
  If I were to continue working on this project, I might consider sub-setting the data by ownership type and looking for correlation and/or difference in strength of relationships among the three main ownership types. I would also consider sub setting the state of Minnesota’s nursing home data and compare it with Wisconsin’s. Population-wise, demographically and economically, the states are similar. Minnesota, however, implemented mandatory statewide reporting of medical errors and misadventures over 15 years ago, with an emphasis already then on transparency. Comparison with Minnesota’s data might help determine if nursing home staff in Wisconsin is reporting incidents.
* Limitations. The data definitions within the metadata document do not specify to the level of where each variable’s data came from. It is entirely possible that the QM Rating variable that I originally planned to use as the outcome variable is actually a derived value – using other explanatory variables from within the same data set. That seems a bit like using your training data to test your model for accuracy!  
  Removal of 7 of the 385 nursing homes due to unsubmitted staffing hours (AKA ‘N/A’s) seemed necessary because I was certain that the cost of paying additional staff would be a key variable in my analysis.
* Concluding Remarks. I was disappointed to find no easy ‘smoking gun’ causality when it came to characteristics that higher performing nursing homes share based on this single data set, but that does not mean that strong correlations do not exist. There are many ways to introduce the concept of quality into a health care data project, and it may be that this particular data set had variables that were too watered-down, as in possibly being derived variables, (or possibly even externally manipulated or gamed) to allow those correlations to be detected given the data I used.