Final Report - Team Uber

1. Introduction

The goal of this project was for our team to analyze data on job postings for two roles within the healthcare vertical and examine what drove job-seeker demand for those roles within the US economy. Ultimately, we wanted to make recommendations to Greenwich, our customer, about how they might sell consulting to other companies using that data, and how to monetize whatever we discovered through our analysis. In the course of this project, we conducted Exploratory Data Analysis, and then pursued a variety of analytical approaches based on that initial exploration. In the end, we identified useful trends relating time to fill a position with the specific tags and variables that were associated with a position, whether it was in an urban or rural area and what region of the country it was in. Lastly, we built a logistic model for predicting whether a particular position would be filled within an acceptable period of time (defined as 30 days or less).

2. Data Cleaning, EDA Examinations

Our goal for data cleaning and exploratory analysis was to identify any data problems that would potentially influence our model building, and provide insights that would guide future analysis. The problems that we addressed through data cleaning were: missing values, and invalid data values. For missing values, there were 222691 and 50728 entries for the nursing assistant and nurse manager roles, respectively. Combined, there were a total of 3.30% of postings which contained missing values for our variable of interest, time_to_fill. Moreover, 4.21% of the postings had missing values in location variables (state, region_state), which would also cause potential analytical problems since a major part of our analysis depended on the location of job postings. Since missing values were only present in approximately 7% of the total data, we removed any entries with missing values in the time_to_fill or region_state variable for future analysis. In the interest of eliminating data values that likely wouldn’t provide value to our analysis, we trimmed the data based on postings with time_to_fill in the lowest and highest quantiles of the data set. We reasoned that positions that filled in an extremely short amount of time or in an unusually long amount of time were likely either postings that were not filled by an outside hire (in the case of short fills) or were not tied to a specific position (in the case of long fills). The final dataset was 85.84% of the original dataset (two roles combined).

In the effort to define demand, we decided to use the time it took to fill a certain job as the indicator for the demand for a job in the current market. In particular, we were interested to see how different factors could be used to predict time_to_fill (so as to predict demand). We selected variables from the dataset, including salary, time and location, and conducted correlation analysis on their relationship with time to fill. However, even though the correlations were significant, no variables were found to be strongly correlated to time_to_fill (max coefficient < 0.05, p < 0.01).

This counterintuitive result triggered our interest and more in depth analysis was conducted to further investigate how time to fill is influenced by different factors, as is reported in the sections below. These initial results also suggested the need to examine other means of grouping the data for comparisons, which led to two geospatial examinations of the data.

3. Rural vs Urban Analysis

From the dataset, we gained plenty of information on the geographical location of each posting, so we came up with an idea of classifying each posting to rural or urban based on its state and county information. In order to do the rural/urban classification, we gleaned the population information from 2010 U.S. Census and attached the rural, urban, total population of each county to our dataset. Based on the rural/urban classification criteria provided by the United States Census Bureau, we classified a county to be rural if it has above 50% of its population living in rural area. Lastly, we created a binary dummy variable in our dataset to indicate the rural/urban classification result of each posting based on the county it belongs to. From the classification result, 8.32% of the nurse manager postings were coming from rural counties and 12.78% of the nursing assistant postings were from rural counties.

After implementing the classification, we investigated the difference in average time to fill of both rural and urban postings within each state. For nurse manager positions (see figure 10), there were a total of 44 states which had job postings from both rural and urban counties. In 25 states, nurse manager positions were filled faster in rural counties. The top three states where the average time to fill in rural counties was much shorter than the average time to fill in urban counties were Alaska, Alabama and Colorado. In Alaska, the average time required to fill a nurse manager position in rural counties was about 39 days shorter than urban counties. In the remaining 19 states, nurse manager positions were filled faster in urban counties. The largest difference in time to fill happened in South Dakota, where the average time required to fill a nurse manager position in urban counties was about 43 days shorter than its rural counties. For nursing assistant (see figure 8), there were 45 states which had postings from both rural and urban counties. In 26 of them, nursing assistant positions were filled faster in rural counties. South Carolina had the largest negative difference in time to fill between its rural and urban counties. Filling a nursing assistant position in rural counties of South Carolina was on average 27 days faster than its urban counties. On the other hand, Nebraska had the largest positive difference (on average 29 days faster in urban counties) in time to fill between its rural and urban counties.

Next, several two-sample t-tests were implemented to further expand our rural/urban analysis. For nursing assistants, the difference in mean of time to fill between all of the rural and urban counties was statistically significant (0.382, 1.234) with p-value 0.0003 (see figure 9). This implied that generally nursing assistant positions could be filled in slightly shorter time in urban counties nationwide. And, surprisingly, for nursing manager, the difference in mean of time to fill between rural and urban counties was actually not significant (-0.946, 1.262) with p-value 0.779 (see figure 11). One last interesting finding was that the salary of nursing assistant in urban counties was 4.6% higher than the salary of nursing assistant in rural counties, while the salary of nurse manager in urban counties was 8.1% higher than the salary in rural counties. Thus, in terms of salary, the rural/urban factor made a larger impact to nurse manager positions.

4. Salary, Number of Posts, Time of the Year (Month) vs Time to Fill

Correlation analysis was conducted to identify potential important factors (time, location, role, number of current postings) that may have a significant relationship with time to fill. Since counterintuitively, no significant relationship was found between salary and time to fill in the exploratory data analysis, we subsetted the data to study the effect of time and location on the relationship between salary and time to fill. Observations were partitioned into 24 different time groups based on the year and month each posting was released (one for each month over two years), and into 51 state groups based on the location of each posting. Whereas the relationship between salary and time to fill remained significant but not strong (max coefficient = 0.067, $p < 0.01$) when partitioning by posting year and month only, the correlations became quite high (max coefficient = 0.909, $p < 0.01$) when grouped by state and posting month together, both for nurse managers and for nurse assistants (see figure 1). This result suggested that there existed a great regional and time difference in terms of the relationship between salary and time-to-fill, and it would be helpful to dissect by region and time of the posting for further analysis on time to fill. In terms of roles, the relationships between salary and time-to-fill were not identical for managers and assistants (correlation coefficient = -0.02, $p = 0.40$), suggesting that different levels in the industry might have different peak times for high demand, and should be analyzed separately.

Moreover, the number of listed postings in a certain state given a certain time was also found to have a significant relationship with time to fill. Since the number of available postings in a market may influence the time it takes to fill vacant jobs, we created a new variable “count” that counted the number of posted jobs by state, post year and post month to study its effect on time to fill, and its interactions with location and time. Through treemaps and correlation analysis, we found that larger states (west and east coast) tended to have more postings in general (not surprisingly), and tended to take longer to fill jobs than middle-sized states (see figures 2 and 3). States with more postings also had stronger, and more positive correlation between salary and time to fill, which suggested that when there were more postings, the time it took for a high salary job to be filled was also longer. This result suggested that the number of jobs posted may have an intermediate effect on the effect of time and location to time to fill, and it would be helpful to put it in the regression model to further test its effect. The correlation visualization and treemaps for this section are shown in the appendix.

5. Tags Analysis

Although no single tag was strongly correlated with the time_to_fill for a posting, we noticed during EDA that certain tags seemed to refer to different categories regarding a job posting. For instance, there were tags that specifically mentioned a level of education, length of experience, or a particular shift. We decided to go through the tags and group them by categories, then examine the trends for time_to_fill for different postings related to each tag within a category. Due to the nature of the data, we were left to infer what specifically a tag referred to, which introduced an element of human error into this analysis, but often it was reasonably clear what a tag was referring to.

Our first job within this analysis was to construct groups of tags and do some additional data cleaning to help ensure a quality of analysis for those tags that were included. To start, we examined the overall count of each tag as well as how often a posting with that tag was missing the time_to_fill value. We then trimmed our data to reflect only those posting with a tag which wasn't in the bottom 20% quantile of absolute count (approximately 23 instances of the tag at a minimum). We then examined the list of unique tags and construct groups. Our first group included different education levels (GED through PhD), our second related to different levels of experience (1 year through 10 years), our third covered different job locations and departments within a hospital, our fourth covered which shift/hours the role covered, and the last group related to language skills (Spanish, Russian, Vietnamese, etc).

We then aggregated the postings for each tag and calculated the average time to fill for that tag, comparing across each grouping. This yielded noticeable differences for each tag (see figures 6 and 7). Not all differences were statistically significant, but checking pairwise differences within each group, a majority of tags were different to a statistically significant extent from at least a few other tags in their group. The result of this analysis is that we can identify tags that tend to be found with lower time_to_fill. Some of these tags will become part of our "successful" job posting logistic model, but, even on its own, this analysis can help a company to focus their efforts on jobs that would be expected to take longer to fill.

One last tags-related analysis had to do with examining the number of tags that a job had associated with it with the average time to fill. We examined this trend for both roles separately, and discovered that jobs with more tags associated tended to fill more slowly than jobs with only a few associated tags (see figures 4 and 5). There are two possible ways to think about jobs with lots of associated tags. On the one hand, it could be that this posting was for a job that seemed to have lots of separate requirements, which might have resulted in fewer applicants feeling they actually fit the description. On the other hand, this could have been a job posting that was very generic, and perhaps not tailored to a specific position to be filled, and as such might have been kept up for a specified period of time regardless of the applicants that were received. Under either of these scenarios we could expect to see a longer time to fill for those postings, so this is not completely unexpected. The number of tags related to a posting did end up as a predictor in our logistic model.

6. Regional Analysis

Another analytic approach undertaken was examining our data for spatial autocorrelation. A common way to look at the economy is by looking at different regions of the country, and so we examined whether looking at our data through the lens of different regions might help us to compare subsets of data. Using some classical definitions of different US regions (west coast, rust belt, etc), we split the data and then conducted a test of spatial autocorrelation called Moran's Index. Moran's Index measures the similarity of points that are close to one another, with a value of one indicating strong spatial autocorrelation, zero indicating a truly random sample of values, and negative one indicating a perfectly spatially uncorrelated set of data (like the colors on a chess board). If our definition of regions provided useful spatial autocorrelation, we would expect to see a Moran's Index between zero and one.

After establishing our regions, we conducted two tests. First, we used the inverse of the distances between states to compute the standard spatial autocorrelation for the US. This resulted in a statistically significant Moran's Index of 0.07. Next we computed the index again, but this time two elements were given a distance of one if they were in the same region, and zero if they were not. This resulted in a statistically significant Moran's Index of 0.26, which is a marked improvement over the standard index based on physical distance. This suggested that our regional definitions were a valuable means of grouping our data into similar sets.

After verifying that the regions as proposed were spatially autocorrelated, we examined the average time to fill for each region. Comparing these, we found differences between regions for each role, as well as between the two roles (see figure 12). We tested these differences for statistical significance, and about half were found to be significant. As such, we were able to identify which reasons filled faster for which roles. From this, Greenwich would be able to suggest to a company which was seeking to fill roles across the country which areas may need additional emphasis or resources in order to fill a posting in a desired amount of time.

7. Predictive Model for Success/Failure of Job Fill

To synthesize our previous analyses, we created a logistic regression model to predict time to fill above and below a specific threshold. First, we removed outliers based on quantiles in the dataset (only including observations with time to fill between the 10th and 90th quantiles, as mentioned in the EDA section). Then, after separating the data into the two roles, we turned time to fill into a binary variable with "success" representing a time to fill less than the median of the dataset (approximately 30 days), and “failure” representing a time to fill greater than this value. For each of our logistic regression models, we separated the datasets into training and test groups and built an initial model, including all variables we believed to be potential predictors. These included the tf-idf values, rural vs. urban classification, and region aggregation variables we had created.

For the Nurse Manager role (see figure 13), significant predictors included post month, the region aggregate variable, salary, number of tags, and certain tf-idf values(Critical Care, Discharge planning, Computer skills). With an F1-score of 0.64, this logistic regression model adequately predicted the binary response variable. For the Nurse Assistant role (see figure 14), significant predictors included post month, the region aggregate variable, salary, number of tags, and certain tf-idf values (Emergency, Veteran, Work.Weekend, and Assisted.Living). With an F1-score of 0.59, this logistic regression model clearly needed some refinement, but still somewhat adequately predicted the binary response variable.

Significant Predictors: Post Month, Region, salary, #Tags, tf-idf values (Emergency, Veteran, Work.Weekend, Assisted.Living). Summary results of the logistic models are shown in the appendix 9.5 below.

8. Conclusions

Based on our analysis, posting month, region, salary, number of associated tags and tf-idf values including emergency, veteran, work.weekend and assisted.living were all significant in predicting whether a job posting would be filled in a month or not. This result suggested a clear seasonality as well as regional trend in the job demand for nurse manager and assistant. Therefore, in future endeavors, to ensure new postings can be filled within a certain time frame, it would be helpful to release them in the high demand months given the region of the posting. Moreover, since the number of tags was found to be negatively related to whether a job would be filled in time, having less, but more specific, tags for each posting might also be a good way to efficiently market each posting. We believe by tailoring job postings based on time, region, tags, and tf-idf values, as our analysis suggested above, new postings will be able to reach more interested job seekers and as a result be filled more efficiently.

9. Appendix

9.1 Correlation between salary and time_to_fill for each state for each month

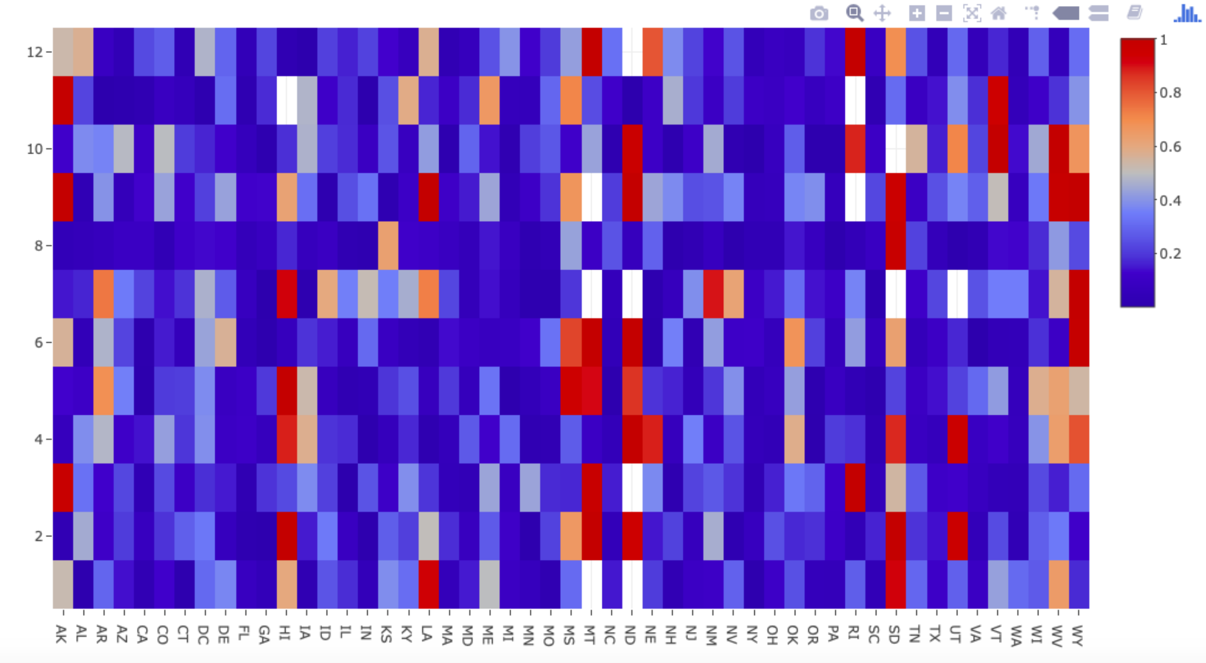


Fig 1. Treemap for number of job posts and the correlation between time_to_fill and salary by state

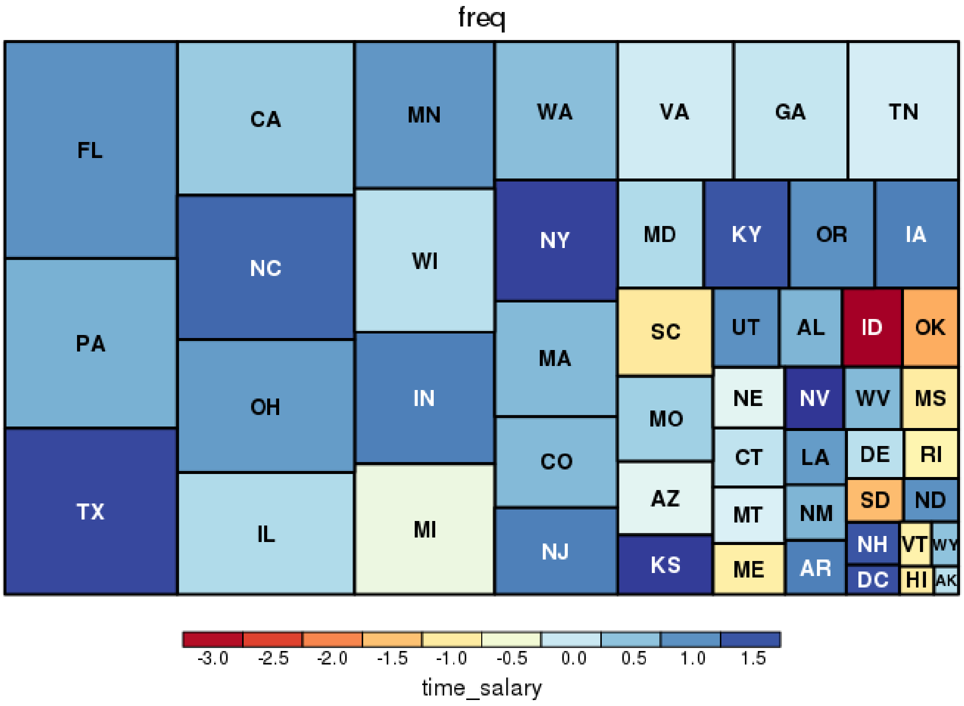


Fig 2. Treemap for number of job posts and the mean time it takes to fill a job by state

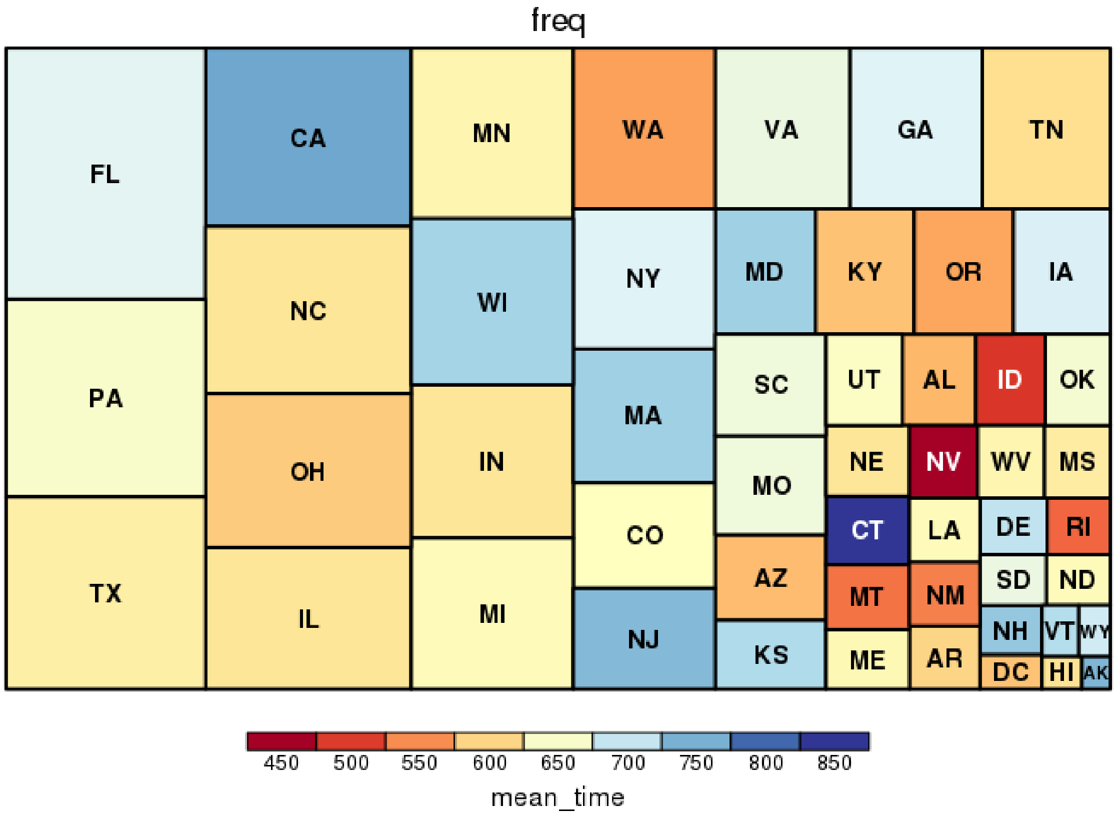


Fig 3. Treemap for the mean time it takes to fill a job by state

9.2 Tags Analysis

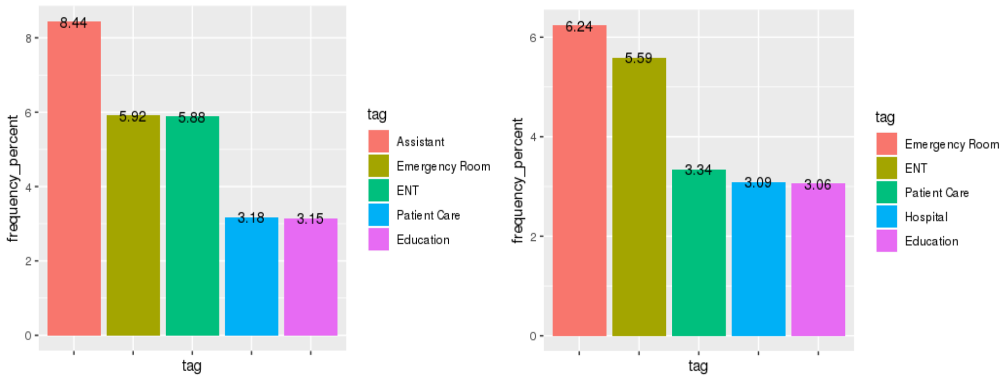


Fig 4. Tag frequency for Nurse Manager (left) and Nursing Assistant(right)

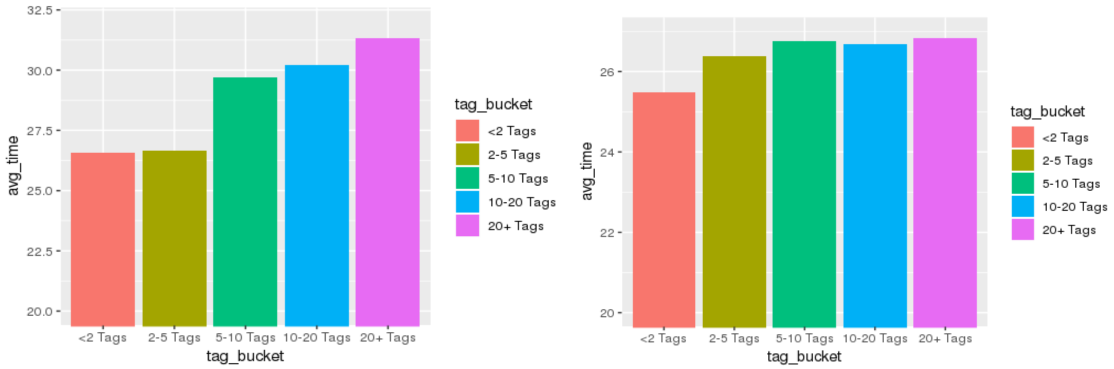


Fig 5. #Tags vs time_to_fill for Nurse Manager (left) and Nursing Assistant(right)

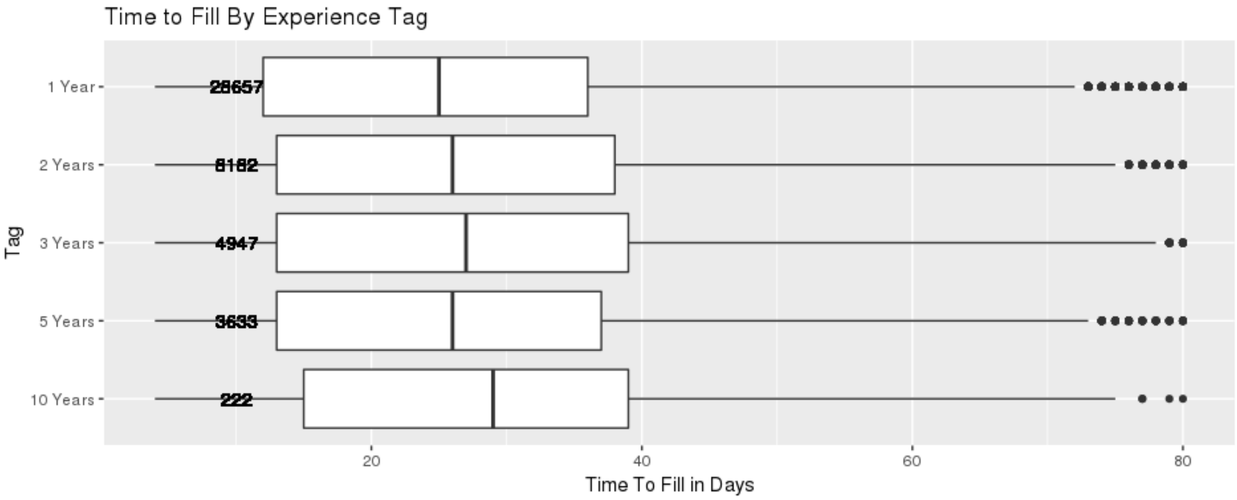


Fig 6. Time to Fill for Various Experience Tags

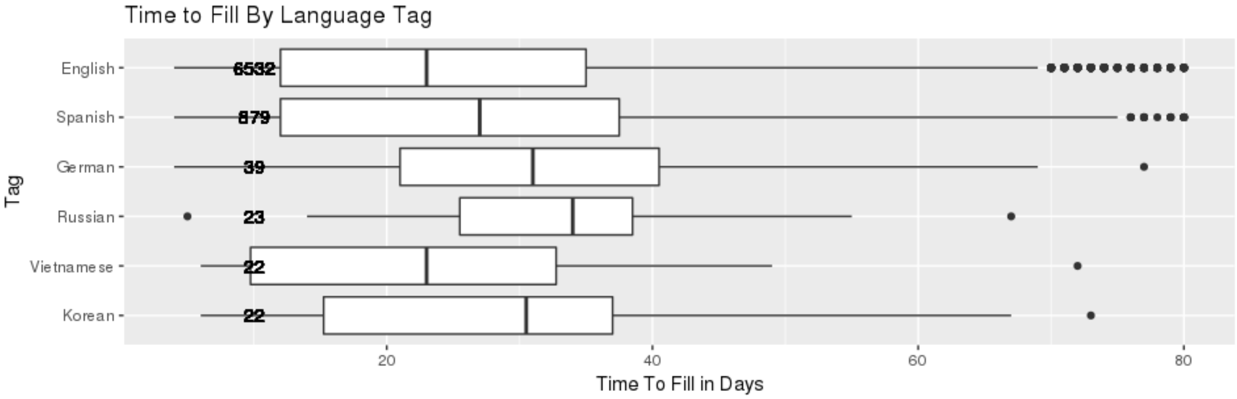
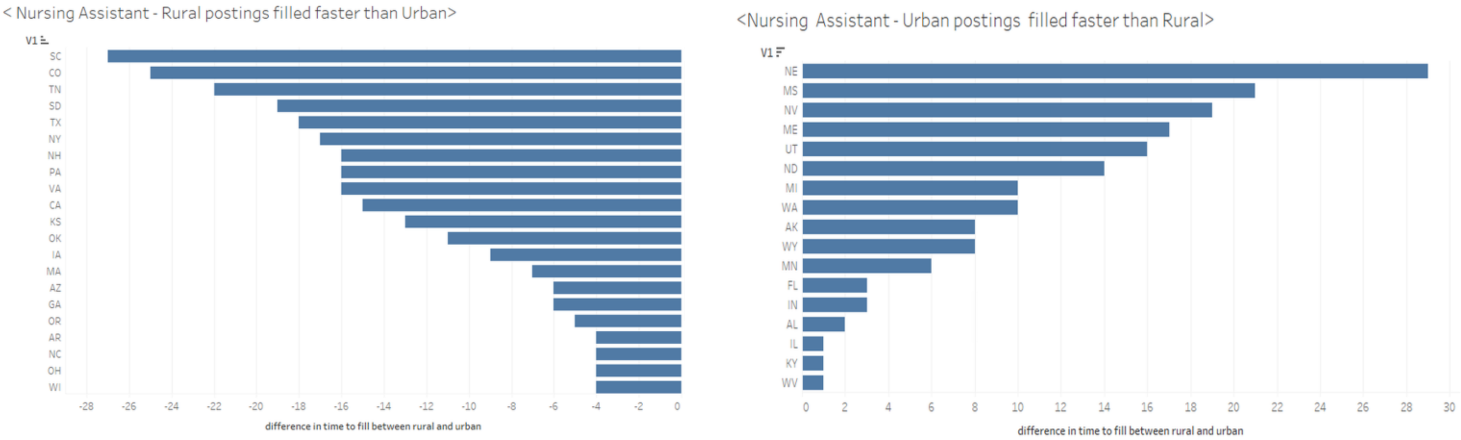


Fig 7. Time to Fill for Various Language Tags

9.3 Rural vs Urban Analysis

- Nursing Assistant


< Nursing Assistant - Rural postings filled faster than Urban>

<Nursing Assistant - Urban postings filled faster than Rural>

Fig 8. States where postings were filled faster in rural counties (Left)
States where postings were filled faster in urban counties(Right)

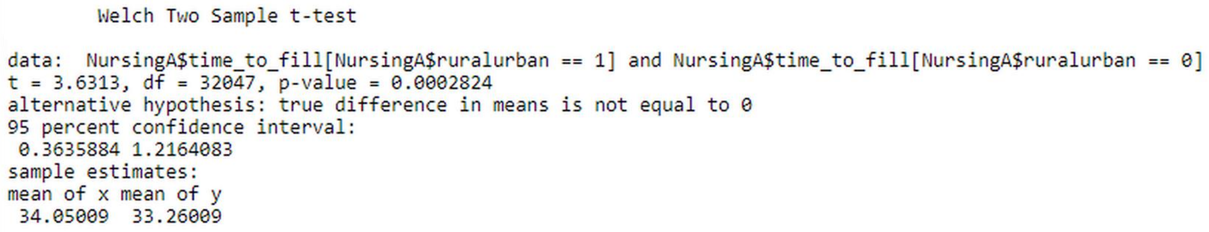
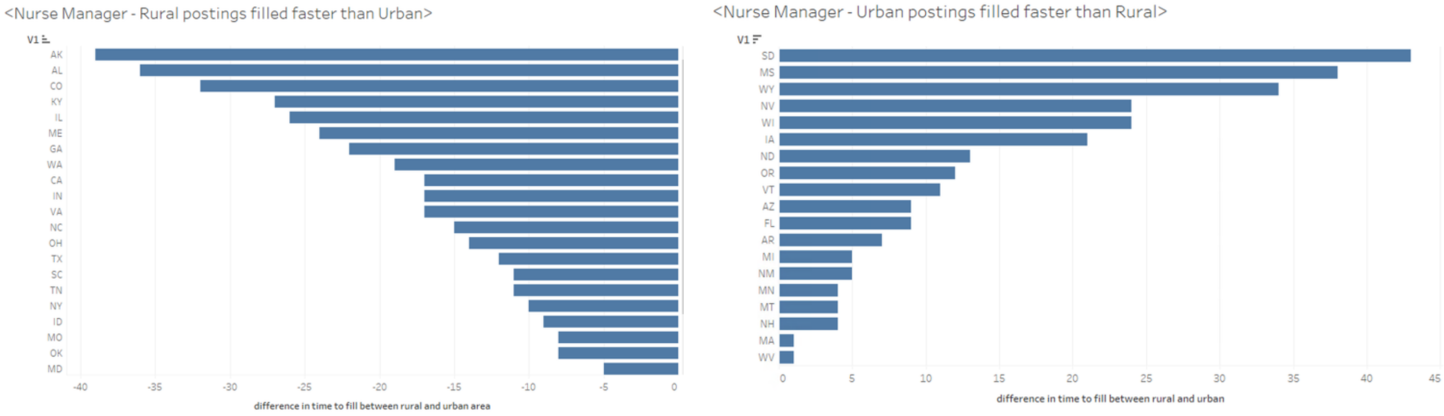

```
        Welch Two Sample t-test

data:  NursingA$time_to_fill[NursingA$ruralurban == 1] and NursingA$time_to_fill[NursingA$ruralurban == 0]
t = 3.6313, df = 32047, p-value = 0.0002824
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 0.3635884 1.2164083
sample estimates:
mean of x mean of y
 34.05009  33.26009
```

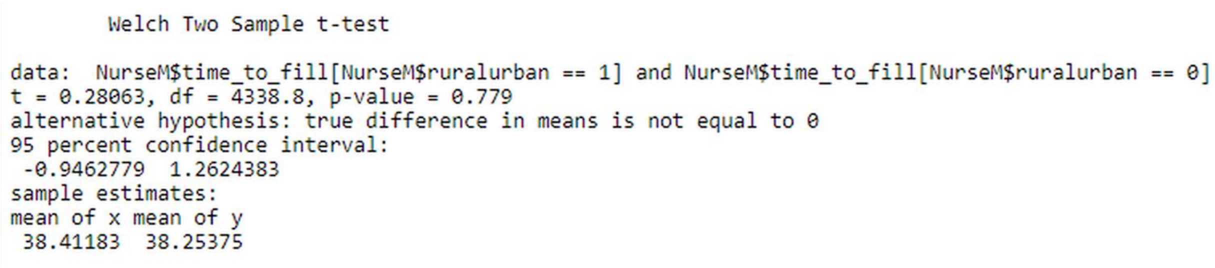
Fig 9. Two sample t-test result for difference in mean of time to fill between rural urban counties

- Nurse Manager


<Nurse Manager - Rural postings filled faster than Urban>

<Nurse Manager - Urban postings filled faster than Rural>

Fig 10. States where postings were filled faster in rural counties (Left)
States where postings were filled faster in urban counties(Right)



```
        Welch Two Sample t-test

data:  NurseM$time_to_fill[NurseM$ruralurban == 1] and NurseM$time_to_fill[NurseM$ruralurban == 0]
t = 0.28063, df = 4338.8, p-value = 0.779
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.9462779  1.2624383
sample estimates:
mean of x mean of y
 38.41183  38.25375
```

Fig 11. Two sample t-test result for difference in mean of time to fill between rural urban counties

9.4 Regional Analysis

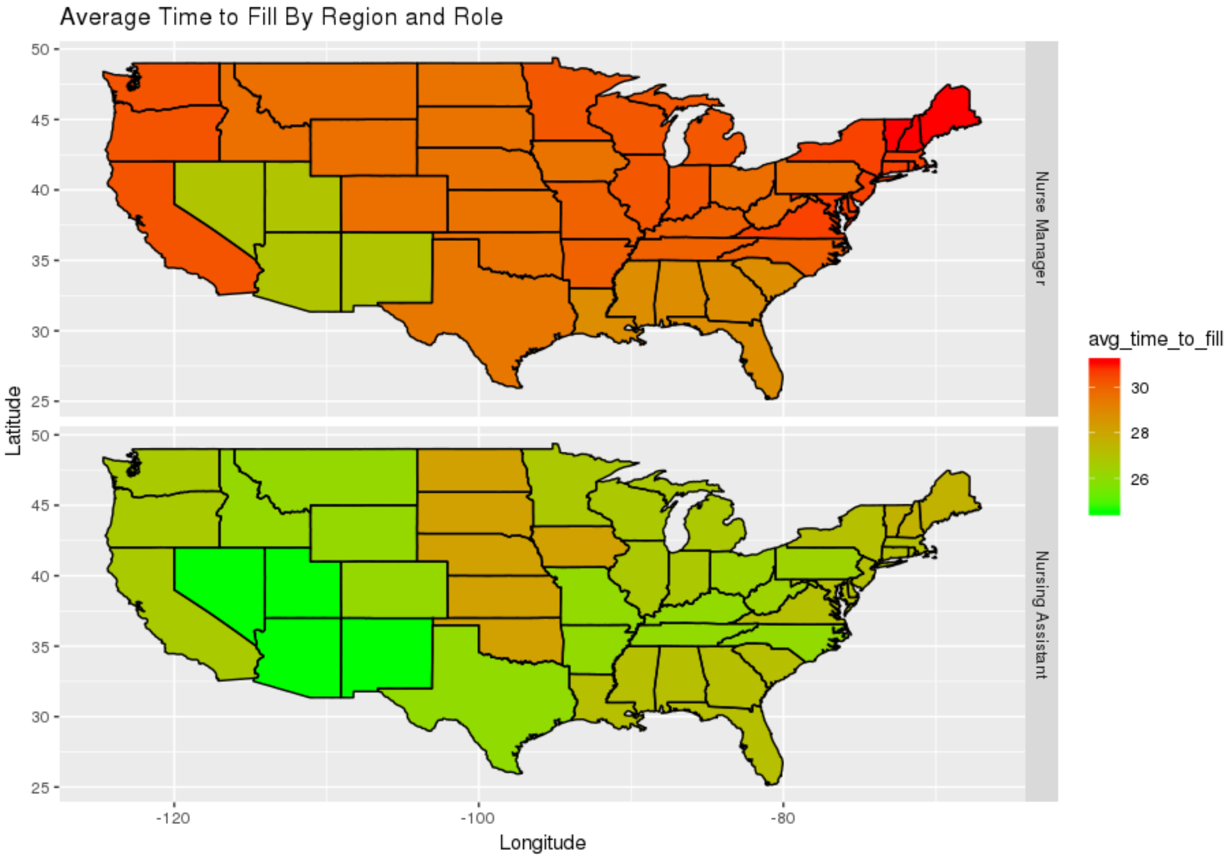
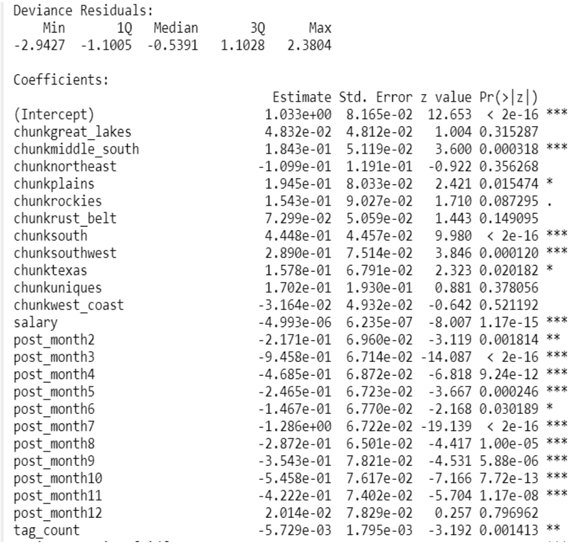


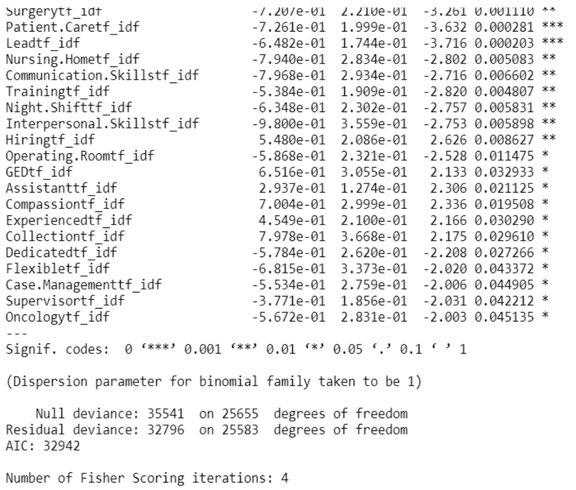
Fig 12. Regional Differences Breakdown for Both Roles

9.5 Logistic Regression Model Results and Statistics

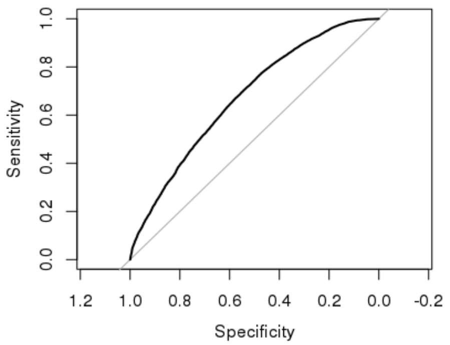
Nurse Manager

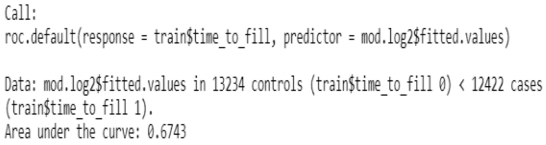


```
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.9427  -1.1005  -0.5391   1.1028   2.3804

Coefficients:
                                  Estimate Std. Error z value Pr(>|z|)
(Intercept)                      1.033e+00  8.165e-02  12.653  < 2e-16 ***
chunkgreat_lakes                 4.832e-02  4.812e-02   1.004 0.315287
chunkmiddle_south                1.843e-01  5.119e-02   3.600 0.000318 ***
chunknortheast                  -1.099e-01  1.191e-01  -0.922 0.356268
chunkplains                      1.945e-01  8.033e-02   2.421 0.015474 *
chunkrockies                     1.543e-01  9.027e-02   1.710 0.087295 .
chunkrust_belt                   7.299e-02  5.059e-02   1.443 0.149095
chunksouth                       4.448e-01  4.457e-02   9.980  < 2e-16 ***
chunksouthwest                   2.890e-01  7.514e-02   3.846 0.000120 ***
chunktexas                       1.578e-01  6.791e-02   2.323 0.020182 *
chunkuniques                     1.702e-01  1.930e-01   0.881 0.378056
chunkwest_coast                 -3.164e-02  4.932e-02  -0.642 0.521192
salary                          -4.993e-06  6.235e-07  -8.007 1.17e-15 ***
post_month2                     -2.171e-01  6.960e-02  -3.119 0.001814 **
post_month3                     -9.458e-01  6.714e-02 -14.087  < 2e-16 ***
post_month4                     -4.685e-01  6.872e-02  -6.818 9.24e-12 ***
post_month5                     -2.465e-01  6.723e-02  -3.667 0.000246 ***
post_month6                     -1.467e-01  6.770e-02  -2.168 0.030189 *
post_month7                     -1.286e+00  6.722e-02 -19.139  < 2e-16 ***
post_month8                     -2.872e-01  6.501e-02  -4.417 1.00e-05 ***
post_month9                     -3.543e-01  7.821e-02  -4.531 5.88e-06 ***
post_month10                    -5.458e-01  7.617e-02  -7.166 7.72e-13 ***
post_month11                    -4.222e-01  7.402e-02  -5.704 1.17e-08 ***
post_month12                     2.014e-02  7.829e-02   0.257 0.796962
tag_count                       -5.729e-03  1.795e-03  -3.192 0.001413 **
```

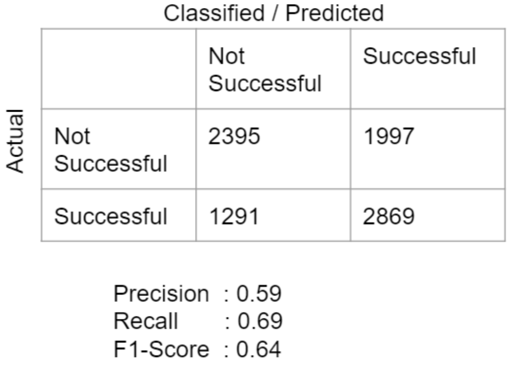


```
Surgerytf_idf                   -7.207e-01  2.210e-01  -3.261 0.001110 **
Patient.Caretf_idf              -7.261e-01  1.999e-01  -3.632 0.000281 ***
Leadtf_idf                      -6.482e-01  1.744e-01  -3.716 0.000203 ***
Nursing.Hometf_idf              -7.940e-01  2.834e-01  -2.802 0.005083 **
Communication.Skillstf_idf      -7.968e-01  2.934e-01  -2.716 0.006602 **
Trainingtf_idf                  -5.384e-01  1.909e-01  -2.820 0.004807 **
Night.Shifttf_idf               -6.348e-01  2.302e-01  -2.757 0.005831 **
Interpersonal.Skillstf_idf      -9.800e-01  3.559e-01  -2.753 0.005898 **
Hiringtf_idf                     5.480e-01  2.086e-01   2.626 0.008627 **
Operating.Roomtf_idf            -5.868e-01  2.321e-01  -2.528 0.011475 *
GEDtf_idf                        6.516e-01  3.055e-01   2.133 0.032933 *
Assistanttf_idf                  2.937e-01  1.274e-01   2.306 0.021125 *
Compassiontf_idf                 7.004e-01  2.999e-01   2.336 0.019508 *
Experiencedtf_idf                4.549e-01  2.100e-01   2.166 0.030290 *
Collectiontf_idf                 7.978e-01  3.668e-01   2.175 0.029610 *
Dedicatedtf_idf                 -5.784e-01  2.620e-01  -2.208 0.027266 *
Flexibletf_idf                  -6.815e-01  3.373e-01  -2.020 0.043372 *
Case.Managementtf_idf           -5.534e-01  2.759e-01  -2.006 0.044905 *
Supervisortf_idf                -3.771e-01  1.856e-01  -2.031 0.042212 *
Oncologytf_idf                  -5.672e-01  2.831e-01  -2.003 0.045135 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 35541  on 25655  degrees of freedom
Residual deviance: 32796  on 25583  degrees of freedom
AIC: 32942

Number of Fisher Scoring iterations: 4
```





```
Call:
roc.default(response = train$time_to_fill, predictor = mod.log2$fitted.values)

Data: mod.log2$fitted.values in 13234 controls (train$time_to_fill 0) < 12422 cases
(train$time_to_fill 1).
Area under the curve: 0.6743
```



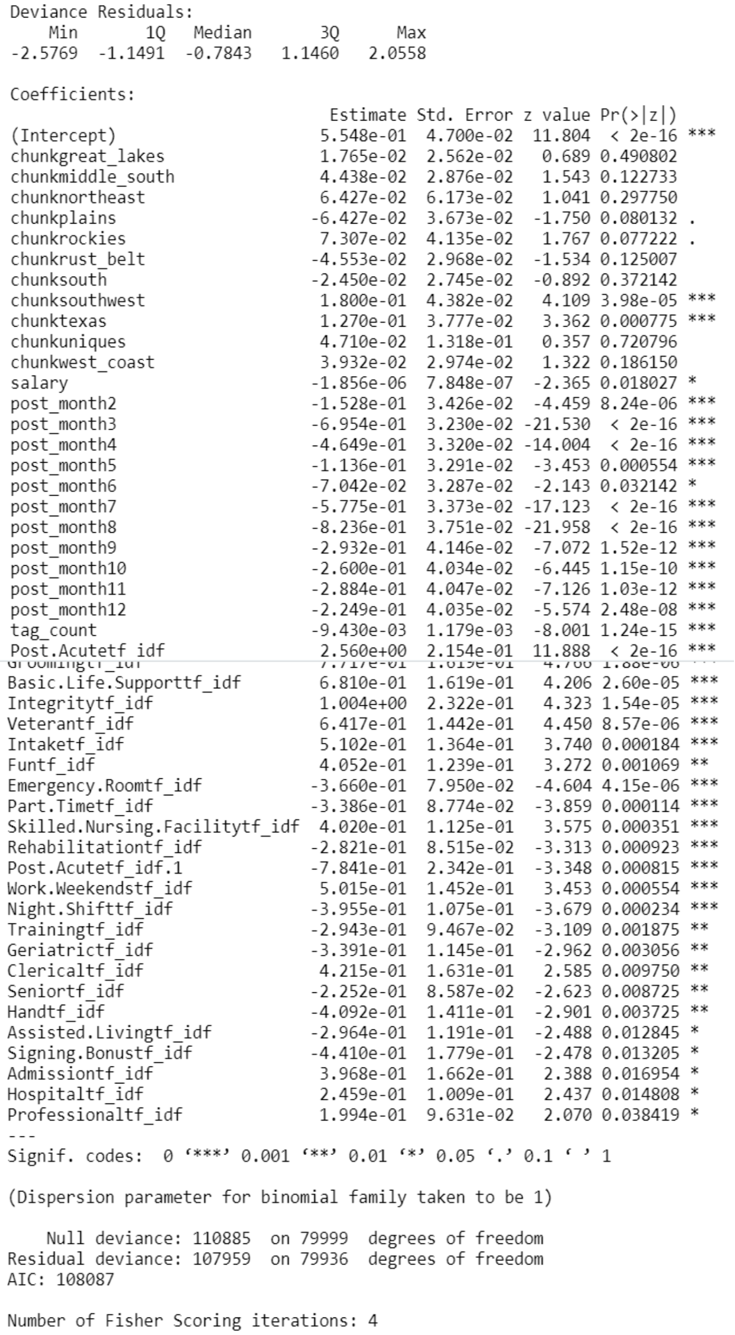
Classified / Predicted

Actual		Not Successful	Successful
	Not Successful	2395	1997
	Successful	1291	2869

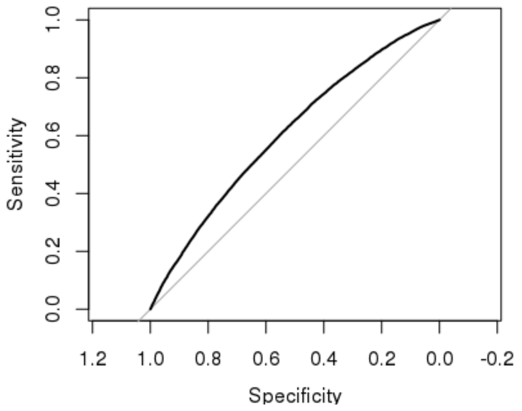
Precision : 0.59
Recall : 0.69
F1-Score : 0.64

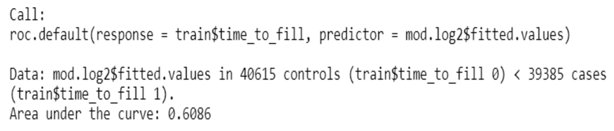
Fig 13. Nurse Manager Logistical Model and Diagnostics

Nursing Assistant

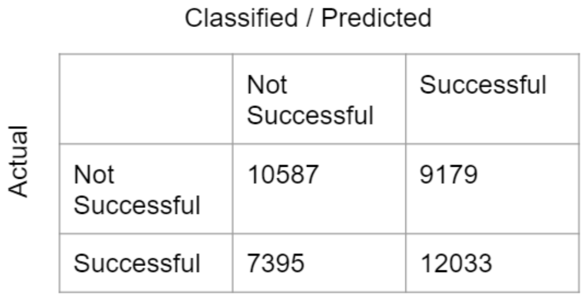


```
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.5769  -1.1491  -0.7843   1.1460   2.0558

Coefficients:
                                 Estimate Std. Error z value Pr(>|z|)
(Intercept)                     5.548e-01  4.700e-02  11.804  < 2e-16 ***
chunkgreat_lakes                1.765e-02  2.562e-02   0.689 0.490802
chunkmiddle_south               4.438e-02  2.876e-02   1.543 0.122733
chunknortheast                  6.427e-02  6.173e-02   1.041 0.297750
chunkplains                    -6.427e-02  3.673e-02  -1.750 0.080132 .
chunkrockies                    7.307e-02  4.135e-02   1.767 0.077222 .
chunkrust_belt                 -4.553e-02  2.968e-02  -1.534 0.125007
chunksouth                     -2.450e-02  2.745e-02  -0.892 0.372142
chunksouthwest                  1.800e-01  4.382e-02   4.109 3.98e-05 ***
chunktexas                      1.270e-01  3.777e-02   3.362 0.000775 ***
chunkuniques                    4.710e-02  1.318e-01   0.357 0.720796
chunkwest_coast                 3.932e-02  2.974e-02   1.322 0.186150
salary                         -1.856e-06  7.848e-07  -2.365 0.018027 *
post_month2                    -1.528e-01  3.426e-02  -4.459 8.24e-06 ***
post_month3                    -6.954e-01  3.230e-02 -21.530  < 2e-16 ***
post_month4                    -4.649e-01  3.320e-02 -14.004  < 2e-16 ***
post_month5                    -1.136e-01  3.291e-02  -3.453 0.000554 ***
post_month6                    -7.042e-02  3.287e-02  -2.143 0.032142 *
post_month7                    -5.775e-01  3.373e-02 -17.123  < 2e-16 ***
post_month8                    -8.236e-01  3.751e-02 -21.958  < 2e-16 ***
post_month9                    -2.932e-01  4.146e-02  -7.072 1.52e-12 ***
post_month10                   -2.600e-01  4.034e-02  -6.445 1.15e-10 ***
post_month11                   -2.884e-01  4.047e-02  -7.126 1.03e-12 ***
post_month12                   -2.249e-01  4.035e-02  -5.574 2.48e-08 ***
tag_count                      -9.430e-03  1.179e-03  -8.001 1.24e-15 ***
Post.Acutetf_idf                2.560e+00  2.154e-01  11.888  < 2e-16 ***
Groomingtf_idf                  7.717e-01  1.619e-01   4.766 1.88e-06 ***
Basic.Life.Supporttf_idf        6.810e-01  1.619e-01   4.206 2.60e-05 ***
Integritytf_idf                 1.004e+00  2.322e-01   4.323 1.54e-05 ***
Veterantf_idf                   6.417e-01  1.442e-01   4.450 8.57e-06 ***
Intaketf_idf                    5.102e-01  1.364e-01   3.740 0.000184 ***
Funtf_idf                       4.052e-01  1.239e-01   3.272 0.001069 **
Emergency.Roomtf_idf           -3.660e-01  7.950e-02  -4.604 4.15e-06 ***
Part.Timetf_idf                -3.386e-01  8.774e-02  -3.859 0.000114 ***
Skilled.Nursing.Facilitytf_idf  4.020e-01  1.125e-01   3.575 0.000351 ***
Rehabilitationtf_idf           -2.821e-01  8.515e-02  -3.313 0.000923 ***
Post.Acutetf_idf.1             -7.841e-01  2.342e-01  -3.348 0.000815 ***
Work.Weekendstf_idf             5.015e-01  1.452e-01   3.453 0.000554 ***
Night.Shifttf_idf              -3.955e-01  1.075e-01  -3.679 0.000234 ***
Trainingtf_idf                 -2.943e-01  9.467e-02  -3.109 0.001875 **
Geriatrictf_idf                -3.391e-01  1.145e-01  -2.962 0.003056 **
Clericaltf_idf                  4.215e-01  1.631e-01   2.585 0.009750 **
Seniortf_idf                   -2.252e-01  8.587e-02  -2.623 0.008725 **
Handtf_idf                     -4.092e-01  1.411e-01  -2.901 0.003725 **
Assisted.Livingtf_idf          -2.964e-01  1.191e-01  -2.488 0.012845 *
Signing.Bonustf_idf            -4.410e-01  1.779e-01  -2.478 0.013205 *
Admissiontf_idf                 3.968e-01  1.662e-01   2.388 0.016954 *
Hospitaltf_idf                  2.459e-01  1.009e-01   2.437 0.014808 *
Professionaltf_idf              1.994e-01  9.631e-02   2.070 0.038419 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 110885  on 79999  degrees of freedom
Residual deviance: 107959  on 79936  degrees of freedom
AIC: 108087

Number of Fisher Scoring iterations: 4
```





```
Call:
roc.default(response = train$time_to_fill, predictor = mod.log2$fitted.values)

Data: mod.log2$fitted.values in 40615 controls (train$time_to_fill 0) < 39385 cases
(train$time_to_fill 1).
Area under the curve: 0.6086
```



Classified / Predicted

Actual		Not Successful	Successful
	Not Successful	10587	9179
	Successful	7395	12033

Precision : 0.57
Recall : 0.62
F1-Score : 0.59

Fig 14. Nursing Assistant Logistical Model and Diagnostics

10. References

- Rural, Urban population data: https://www.census.gov/data/datasets/2017/demo/popest/counties-total.html

- Rural/Urban classification criterion: https://www.census.gov/geo/reference/urban-rural.html