# RA Task Report

## 1. Introduction

This report summarizes key findings and insights derived from the analysis of the provided data. In section 2, I will briefly review the data wrangling processes that were performed to prepare the data. In section 3, I will discuss the key findings and will address the questions. At the end, I will propose further steps to be done in the future.

Tools used in the analysis includes Python's widely used libraries such as Pandas, Numpy, Matplotlib, and Seaborn. Pandas provides an efficient framework for the analysis of data, and it integrates well with other libraries.

## 2. Data Wrangling

Before performing the analysis, the data was processed. I began by transforming patient's arrival and discharge dates into the datetime data type. Also, I standardized shiftid column (e.g. changed "noon" to "12:00 PM" wherever necessary.) Then I created and transformed the start and the end of shifts based off shiftid column. In addition, I adjusted shift_end date in cases where it rolls over the next day.

## 3. Interpretation and Findings

### 3.1. Summary Statistics and Potential Outliers

Several types of potential data entry errors were found throughout different data quality checks. The presence of outliers in the data could potentially suggest data entry error. As indicated in the below tables, some patients waited around 5 hours to be seen by a physician, which seems unrealistic. Likewise, some physicians had to work more than a day past the end of their shift. It's worth talking to data engineers to find the sources of these errors.

The list of various cases follows:

- Lengths of stay that are too short or too long
- Lengths of stay that are negative

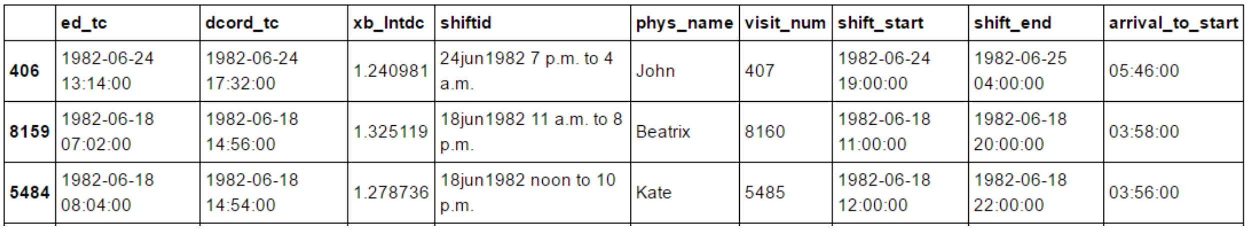
The threshold for outlier detection was defined as:

$$High = Q3 + IQR$$

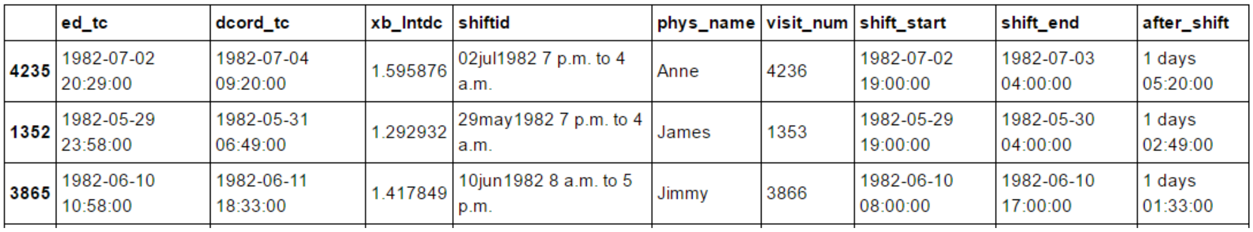
$$Low = Q1 - IQR$$

Where Q1, Q3, IQR are the first quartile, the third quartile, and the interquartile range, respectively.

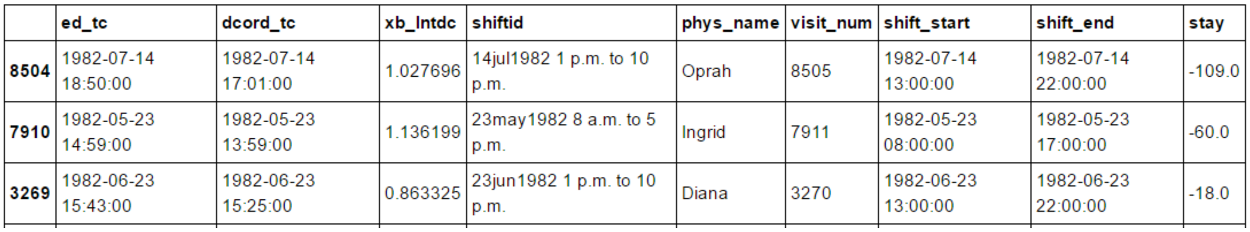
For the length of stay, I assumed lengths shorter than 5 minutes (including negative values) to be outliers.



| | ed_tc | dcord_tc | xb_lntdc | shiftid | phys_name | visit_num | shift_start | shift_end | arrival_to_start |
|---|---|---|---|---|---|---|---|---|---|
| **406** | 1982-06-24 13:14:00 | 1982-06-24 17:32:00 | 1.240981 | 24jun1982 7 p.m. to 4 a.m. | John | 407 | 1982-06-24 19:00:00 | 1982-06-25 04:00:00 | 05:46:00 |
| **8159** | 1982-06-18 07:02:00 | 1982-06-18 14:56:00 | 1.325119 | 18jun1982 11 a.m. to 8 p.m. | Beatrix | 8160 | 1982-06-18 11:00:00 | 1982-06-18 20:00:00 | 03:58:00 |
| **5484** | 1982-06-18 08:04:00 | 1982-06-18 14:54:00 | 1.278736 | 18jun1982 noon to 10 p.m. | Kate | 5485 | 1982-06-18 12:00:00 | 1982-06-18 22:00:00 | 03:56:00 |



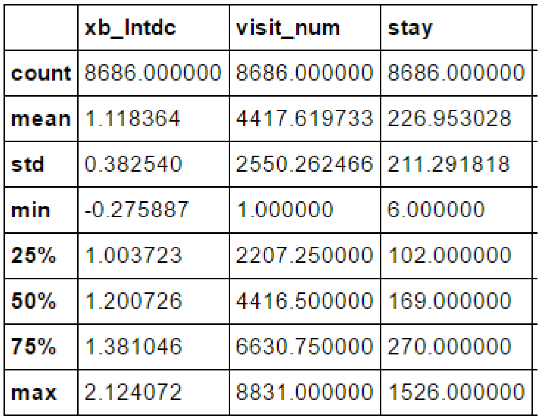
| | ed_tc | dcord_tc | xb_lntdc | shiftid | phys_name | visit_num | shift_start | shift_end | after_shift |
|---|---|---|---|---|---|---|---|---|---|
| **4235** | 1982-07-02 20:29:00 | 1982-07-04 09:20:00 | 1.595876 | 02jul1982 7 p.m. to 4 a.m. | Anne | 4236 | 1982-07-02 19:00:00 | 1982-07-03 04:00:00 | 1 days 05:20:00 |
| **1352** | 1982-05-29 23:58:00 | 1982-05-31 06:49:00 | 1.292932 | 29may1982 7 p.m. to 4 a.m. | James | 1353 | 1982-05-29 19:00:00 | 1982-05-30 04:00:00 | 1 days 02:49:00 |
| **3865** | 1982-06-10 10:58:00 | 1982-06-11 18:33:00 | 1.417849 | 10jun1982 8 a.m. to 5 p.m. | Jimmy | 3866 | 1982-06-10 08:00:00 | 1982-06-10 17:00:00 | 1 days 01:33:00 |



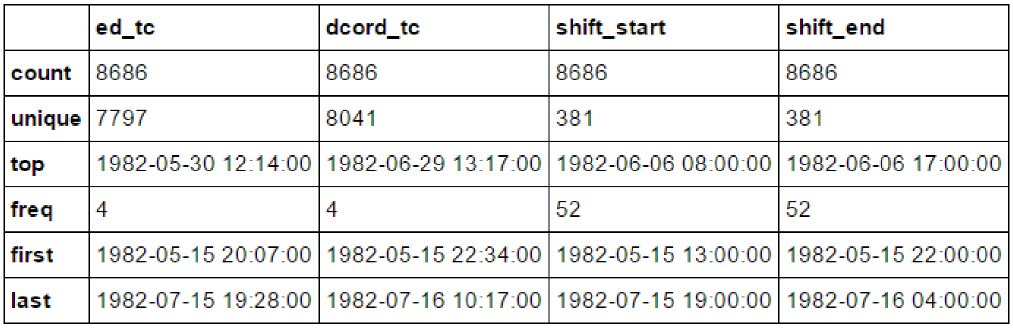
| | ed_tc | dcord_tc | xb_lntdc | shiftid | phys_name | visit_num | shift_start | shift_end | stay |
|---|---|---|---|---|---|---|---|---|---|
| **8504** | 1982-07-14 18:50:00 | 1982-07-14 17:01:00 | 1.027696 | 14jul1982 1 p.m. to 10 p.m. | Oprah | 8505 | 1982-07-14 13:00:00 | 1982-07-14 22:00:00 | -109.0 |
| **7910** | 1982-05-23 14:59:00 | 1982-05-23 13:59:00 | 1.136199 | 23may1982 8 a.m. to 5 p.m. | Ingrid | 7911 | 1982-05-23 08:00:00 | 1982-05-23 17:00:00 | -60.0 |
| **3269** | 1982-06-23 15:43:00 | 1982-06-23 15:25:00 | 0.863325 | 23jun1982 1 p.m. to 10 p.m. | Diana | 3270 | 1982-06-23 13:00:00 | 1982-06-23 22:00:00 | -18.0 |

A total of 138 outliers were removed from the dataframe. It is wise to consult with a subject matter expert to ensure the accuracy of thresholds.

Here is the data summary after the removal of outliers.



| | xb_lntdc | visit_num | stay |
|---|---|---|---|
| **count** | 8686.000000 | 8686.000000 | 8686.000000 |
| **mean** | 1.118364 | 4417.619733 | 226.953028 |
| **std** | 0.382540 | 2550.262466 | 211.291818 |
| **min** | -0.275887 | 1.000000 | 6.000000 |
| **25%** | 1.003723 | 2207.250000 | 102.000000 |
| **50%** | 1.200726 | 4416.500000 | 169.000000 |
| **75%** | 1.381046 | 6630.750000 | 270.000000 |
| **max** | 2.124072 | 8831.000000 | 1526.000000 |



| | ed_tc | dcord_tc | shift_start | shift_end |
|---|---|---|---|---|
| **count** | 8686 | 8686 | 8686 | 8686 |
| **unique** | 7797 | 8041 | 381 | 381 |
| **top** | 1982-05-30 12:14:00 | 1982-06-29 13:17:00 | 1982-06-06 08:00:00 | 1982-06-06 17:00:00 |
| **freq** | 4 | 4 | 52 | 52 |
| **first** | 1982-05-15 20:07:00 | 1982-05-15 22:34:00 | 1982-05-15 13:00:00 | 1982-05-15 22:00:00 |
| **last** | 1982-07-15 19:28:00 | 1982-07-16 10:17:00 | 1982-07-15 19:00:00 | 1982-07-16 04:00:00 |

### 3.2. The Arrive-before-shift Phenomenon

As shown in the code, 7.4 % of patients had to wait until their physicians' shift started. This value was reduced to 7% after the removal of outliers

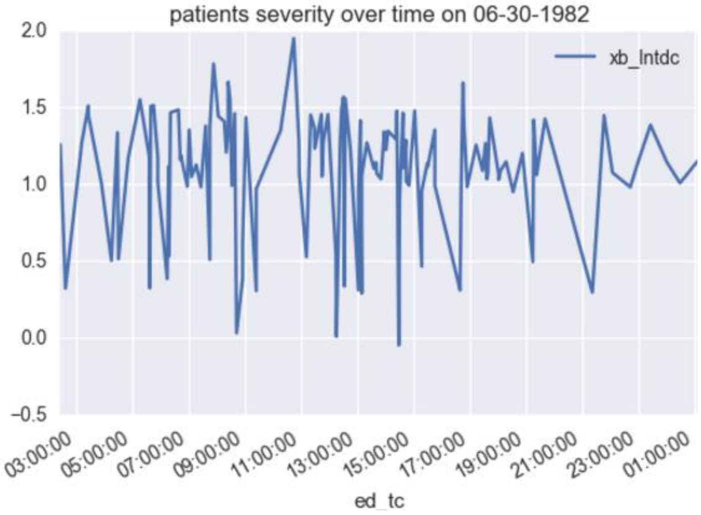
### 3.3. The Discharge-after-shift Phenomenon

As shown in the code, in about 19% of the cases, the physician had to stay past the end of her/his shift. This value was reduced to 17.8% after the removal of outliers

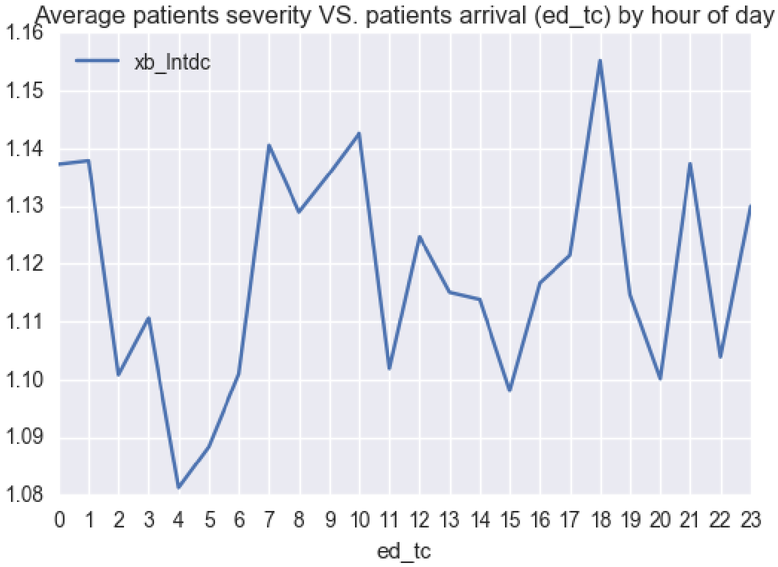
### 3.4. Patterns of Patient's Arrival

Although the pattern seems cyclic, I wouldn't make further interpretation without consulting with a subject matter expert.

The following graph illustrate the trends of patient's severity on an example day (06-30-1982.)

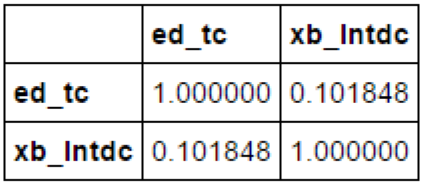


The following graph shows the overall pattern of patients severity by the hour of arrival.



As shown above, there does not seem to be a regular pattern. Also, there is not much consistency between the two graphs. For example, there is a huge decrease in severity after mid-night until 4 AM when we look at the overall pattern, a trend we do not see on the graph corresponding to 06-30-1982. Thus, one can make the conclusion that there are other contributing factor affecting the patient's severity. In section 3.6 we will investigate those factors.

To formally test whether the patient severity is predicted by the hour of the day, one could use Pearson's r. The following table shows there is not a significant correlation between the two variables, however, more investigation is needed to fully understand the nature of these variables.



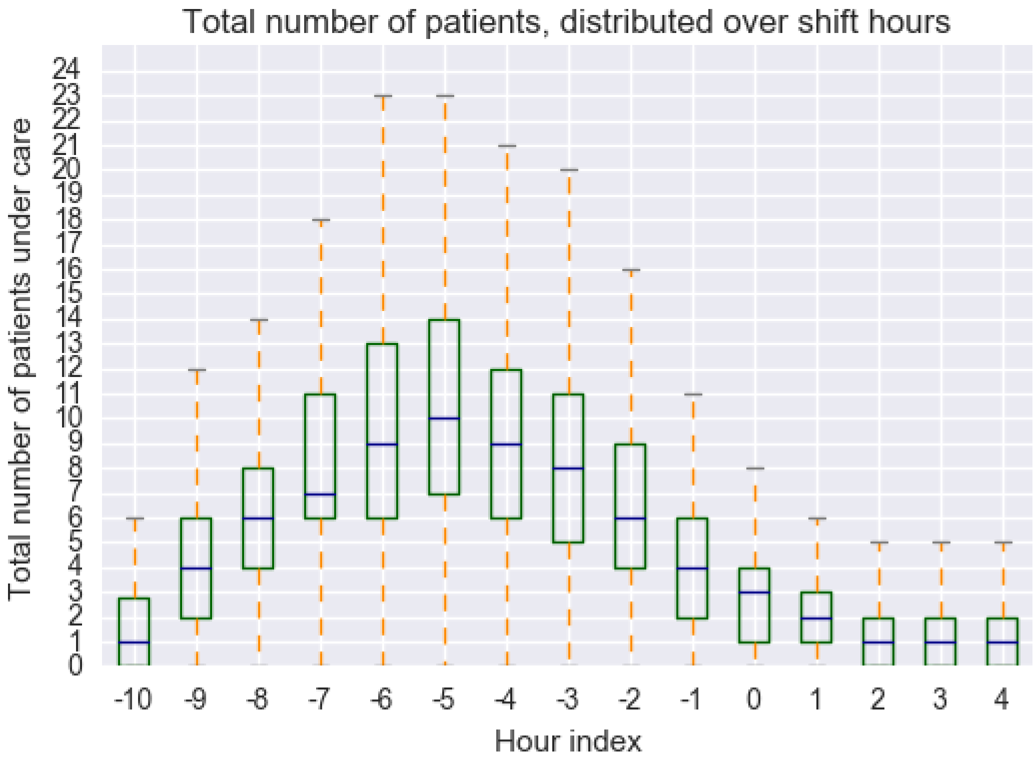
| | ed_tc | xb_lntdc |
|---|---|---|
| **ed_tc** | 1.000000 | 0.101848 |
| **xb_lntdc** | 0.101848 | 1.000000 |

### 3.5. The "Census"

In this section, we create a census data-set as described in part 3 of the questions.

A function was written to create the census. It loops through indices, and for each index that falls into the patient's stay period, it increments the index's value by one. This function uses vectorized operations to avoid expensive loops.

Next, I created a boxplot to demonstrate the distribution of the number of patients under care over indices.



As shown above, the total number of patients under care increases at the beginning of the shifts, it reaches the maximum in the middle of the shift, and it decreases from thereafter.

One limitation of discrete time is that it condenses the variation throughout the hour into one variable. For example, if someone is admitted at 14:40 and is discharged at 15:01, both indices corresponding to 14:00 and 15:00 would increment. The same would happen if those times were 14:00 and 15:59. This may cause inaccuracy when analyzing data.

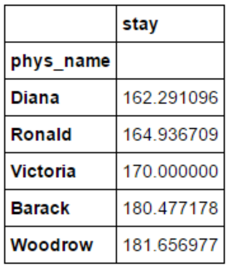
### 3.6. The Fastest Physician

In this section, I'll address the following questions:

- Which physician appears to be the fastest at discharging patients?
  - Short answer: Woodrow
- Which variables do you control for?
  - Short answer: Patients severity
- What are potential threats (and any evidence of them) to your assessment?
  - Short answer: Other contributing factors. Assuming a normal distribution for the corrected stay (rel_stay)

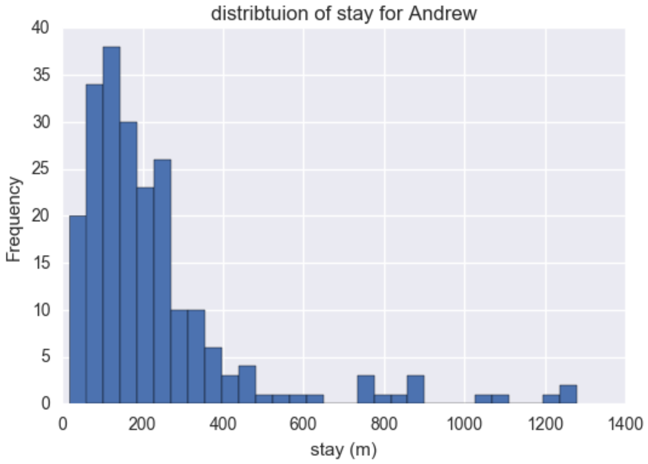
If the distribution of the "stay" variable was normal, we could have just compared physicians based on the mean of stay. But, there are other factors that could have an impact on the variable. We will first look at the distribution for "stay", and then will investigate potential contributing factors.

First, let's look at the mean of the "stay" column for the five fastest physicians.

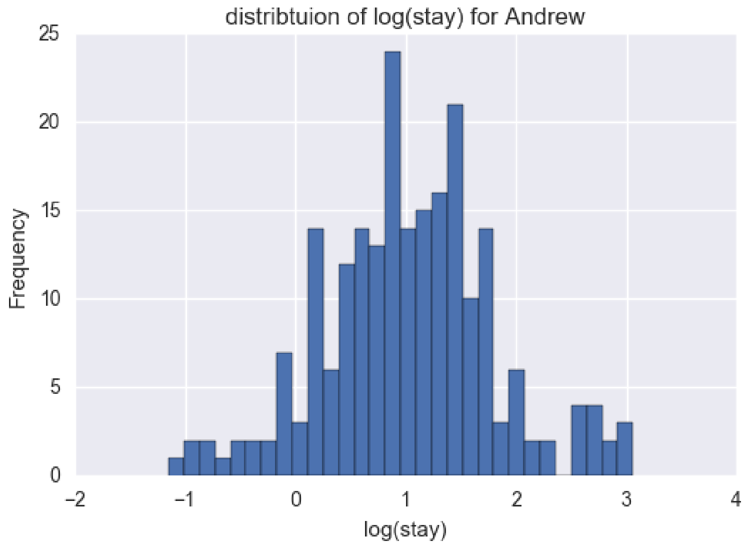


| | **stay** |
|---|---|
| **phys_name** | |
| **Diana** | 162.291096 |
| **Ronald** | 164.936709 |
| **Victoria** | 170.000000 |
| **Barack** | 180.477178 |
| **Woodrow** | 181.656977 |

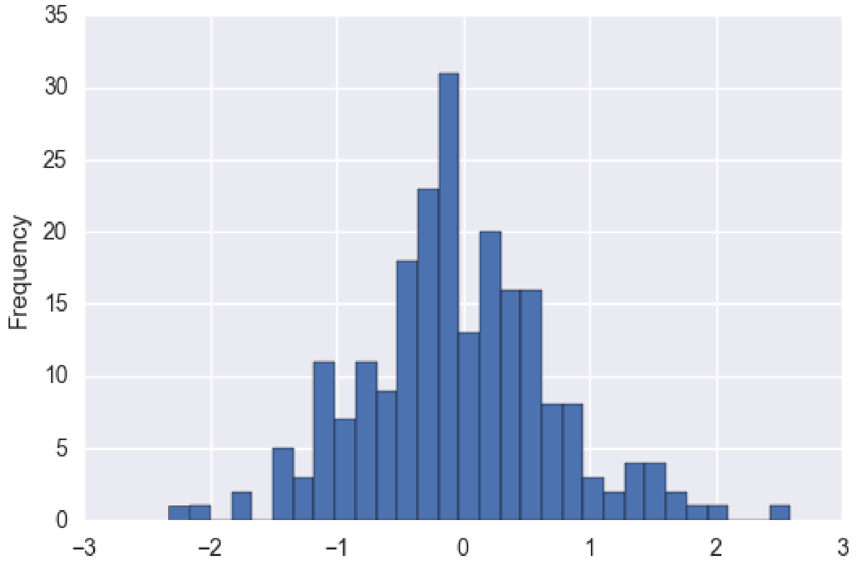
The following graph is the distribution of stay for a physician that was chosen randomly (Andrew).



The above distribution is right skewed. Taking the logarithm of stay may result in a distribution closer to the normal distribution.



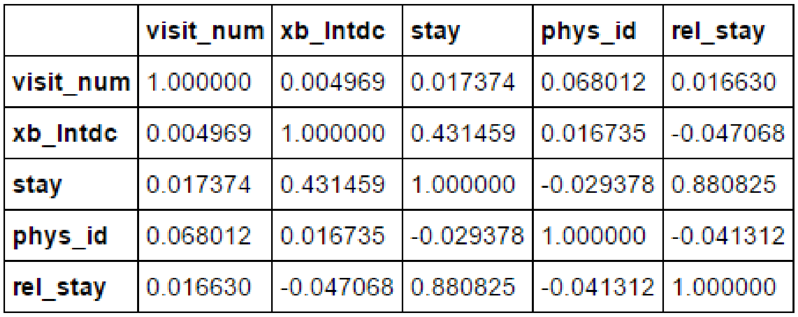
Next, we need to control for the patient's severity. Since xb_lntdc is defined as the logarithm of expected length of stay, we need to treat that as a baseline and see how each data point differs from its corresponding expected value.



We will assume the above distribution is normal, however other factors may play a role:

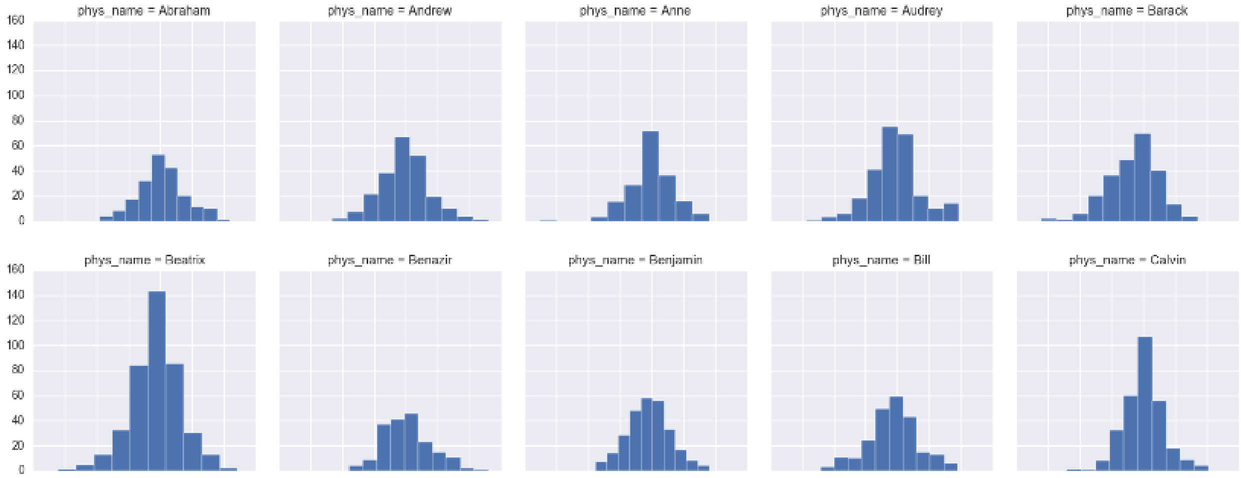
- Number of patients a physician has under care
- Hour of day (We showed that patients severity varies with the hour of the day)
- Hour of shift (One can assume that a physician's performance decreases near the end of shift)

The following table summarizes the correlation between different variables using the Peason’s R.



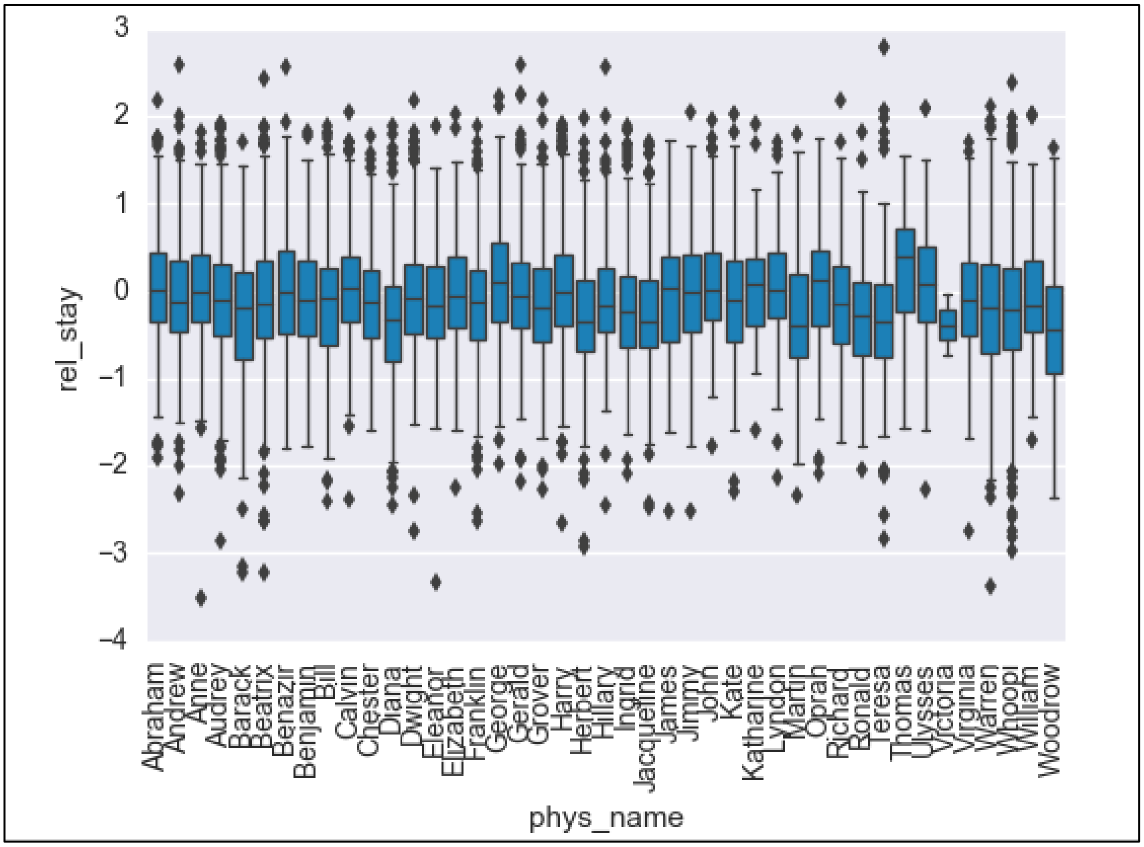
| | visit_num | xb_lntdc | stay | phys_id | rel_stay |
|---|---|---|---|---|---|
| visit_num | 1.000000 | 0.004969 | 0.017374 | 0.068012 | 0.016630 |
| xb_lntdc | 0.004969 | 1.000000 | 0.431459 | 0.016735 | -0.047068 |
| stay | 0.017374 | 0.431459 | 1.000000 | -0.029378 | 0.880825 |
| phys_id | 0.068012 | 0.016735 | -0.029378 | 1.000000 | -0.041312 |
| rel_stay | 0.016630 | -0.047068 | 0.880825 | -0.041312 | 1.000000 |

As shown in the following figure, the distributions of rel_stay look closer to a normal distribution.[1]

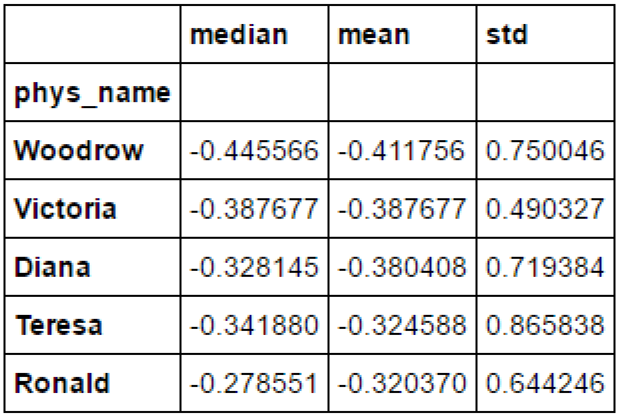


As depicted in the following boxplot, even though physicians' performance fall into a narrow range, it appears that Woodrow is faster than other physicians. Additionally, this graph shows the outliers associated with each physician. This does not include the outliers we removed earlier, but it emphasizes how performances of physicians vary.

[1] Please refer to the python notebook to see the full graph



Finally, below is a list of summary statistics (median, mean, and standard deviations) for the five fastest physicians.



| | median | mean | std |
|---|---|---|---|
| phys_name | | | |
| Woodrow | -0.445566 | -0.411756 | 0.750046 |
| Victoria | -0.387677 | -0.387677 | 0.490327 |
| Diana | -0.328145 | -0.380408 | 0.719384 |
| Teresa | -0.341880 | -0.324588 | 0.865838 |
| Ronald | -0.278551 | -0.320370 | 0.644246 |

## 4. Conclusion

In this report, I explained the key findings discovered by the analysis of the data. While this report provides a basic understanding of the data and the relationship within the variables, more work needs to be done to fully understand the complexity of this data-set.

More work could be done to investigate the nature of correlation between patient's severity and the time under care. A list of the five fastest physicians was provided in section 3.6. However, we need to take into consideration other factors (such as the ones mentioned in section 3.6) that may play a role. One potential approach is to perform a PCA to discover the inter-relationship of different variables. Our assumption that rel_stay is normally distributed requires more testing.