

TravNur

Improving Healthcare Through Data Mining

Comp 541 Data Mining | Project 3

by Verab Chitchyan, Gabriella de Asis, Emmanoel

Dermkrdichyan, Carson Logston, Ahn (Steven) Nguyen

Purpose

TravNur is a nurse placement agency seeking to predict where nurse shortages will occur in order to send nurses to the appropriate locations. Success is defined as the ability to predict hospital nurse shortages with a 70% accuracy 1 week prior to the demand. This allows the agency enough time to hire nurses and send them to a particular location.

Initial Hypotheses

An increase in confirmed covid patients and suspected covid patients at a location will indicate a need for more nurses at that location.

Description of Dataset

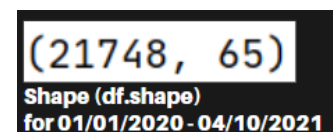
Our dataset has information from hospitals in all 50 states and 3 territories. Data on all 53 locations are reported on a daily basis. Each row represents a state on a given date and has 61 columns of information which report patient statistics, number of occupied beds, number of hospitals facing staff shortages, along with other information.

Our dataset has data types of float64 and object. We had several columns with mixed types which caused a data type warning.

DtypeWarning: Columns (18,20,26,28,30,36,40,56,57,60,61) have mixed types.

state	object
date	object
total number of hospitals	float64
number of hospitals participated	float64

For shape, it has over 21,000 rows and 65 columns. When using the information function, we saw that about half of the column had entries in the 21,000 range which makes sense considering the shape of our dataset. However, we did see a big gap with half of the columns only having about 14,000 entries.



```
(21748, 65)
Shape (df.shape)
for 01/01/2020 - 04/10/2021
```

In evaluating our dataset, we found that most entries prior to July 15th were missing a majority of its information. There was no explanation on the website why reports prior to July 15 were incomplete. Our guess is that states were not asked to report on all 61 attributes prior to July 15 and the requested reported information was expanded on July 15.

Since a majority of the information was missing, we decided to remove entries prior to July 15, 2020. Our dataset now only has information from July 15 2020 to April 10.

DESCRIPTION OF NEW DATASET

```
state      object
date       object
total number of hospitals  float64
number of hospitals participated  float64
```

(14310, 62)

Shape (df.shape)
for 07/15/2020 - 04/10/2021

Data Type (df.dtypes)

```
RangeIndex: 14310 entries, 8 to 14389
Data columns (total 62 columns):
#   Column                                     Non-Null Count  Dtype
---  -
0   state                                     14310 non-null  object
1   date                                     14310 non-null  object
2   number of hospitals participated          14310 non-null  int64
3   critical_staffing_shortage_today_yes     14310 non-null  int64
4   critical_staffing_shortage_today_no     14310 non-null  int64
5   critical_staffing_shortage_today_not_reported 14310 non-null  int64
```

Information (df.info())

Looking at our new dataset, we still have type object and float64 but the mix data type warning is no longer there. We converted each of the attributes with object data types into int data types using string manipulation for easier calculations. For shape, we are now 14,310 by 62. The information on this set shows that most columns have 14,310 non-null entries and the lowest has 14,244 non-null entries.

	previous_day_admission_pediatric_covid_suspected_coverage
count	14310.000000
mean	102.069881
std	94.748319
min	0.000000
25%	45.000000
50%	85.000000
75%	129.000000
max	592.000000

In deciding which descriptive statistics to apply to our dataset, we decided to use the describe function since it gave us the count, mean, standard deviation, min, max, 25th percentile, median, and 75th percentile for each column

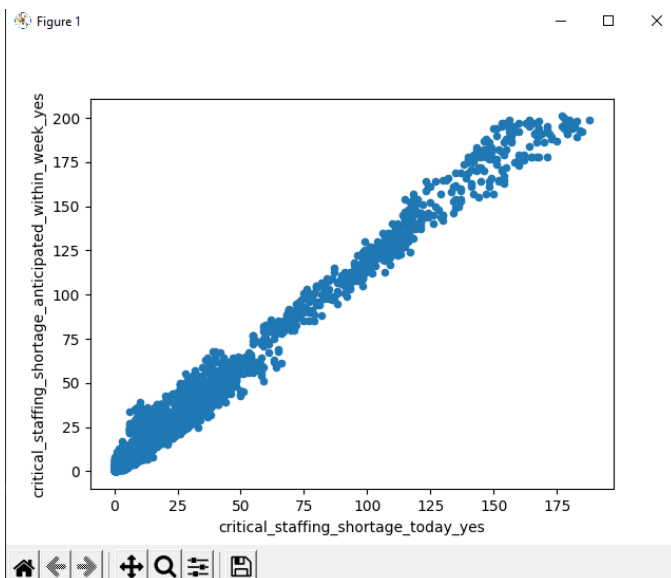
We didn't find it valuable to use the other descriptive statistics such as adding up all the values or the cumulative product of values. The information we found valuable was already displayed using the describe function.

CORRELATIONS

- critical_staffing_shortage_today_yes
- critical_staffing_shortage_anticipated_within_week_yes
- total_adult_patients_hospitalized_confirmed_and_suspected_covid
- total_adult_patients_hospitalized_confirmed_covid
- total_pediatric_patients_hospitalized_confirmed_and_suspected_covid
- total_pediatric_patients_hospitalized_confirmed_covid

total_adult_patients_hospitalized_confirmed_and...	0.813178
total_pediatric_patients_hospitalized_confirmed...	0.813773
total_pediatric_patients_hospitalized_confirmed...	0.813354

Using the correlation function on the columns of our dataset, we found that there was a high correlation between staffing shortages and total number of patients with covid and suspected to have covid. Patients were divided into total adult patients and total pediatric patients. There was also a high correlation between anticipated staffing shortage and patients with covid and suspected to have covid.



We can see that there is a positive correlation between the amount of shortages that occur on a day to day basis and the amount of shortages that are anticipated to occur for that week. The uphill pattern from left to right is the indicator for this positive relationship. We can conclude from this graph that as the number of staff shortages increases, the amount of anticipated staffing required by the hospitals increases as well.

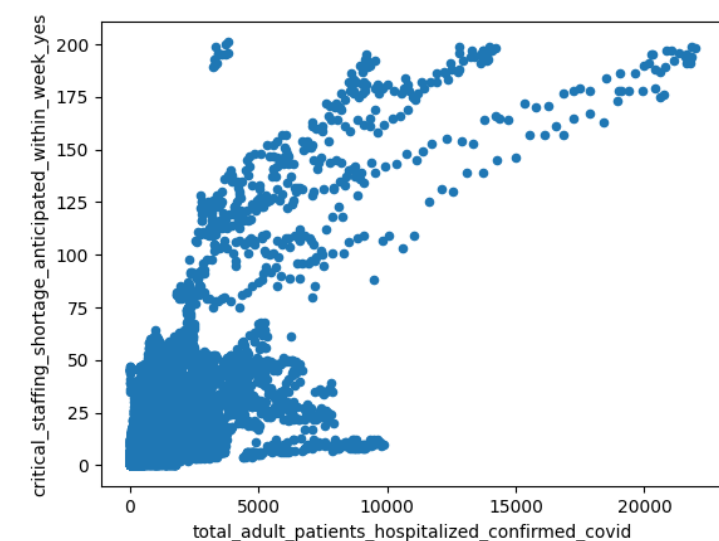
For our company TravNur this relationship demonstrates the needs that we are trying to fulfill, this being the efficiency of getting nurses to hospitals where shortages are clear and defined.

Variables:

- critical_staffing_shortage_anticipated_within_week_ys (y)

critical staffing_shortage_today_yes (x)

Figure 1



Variables:

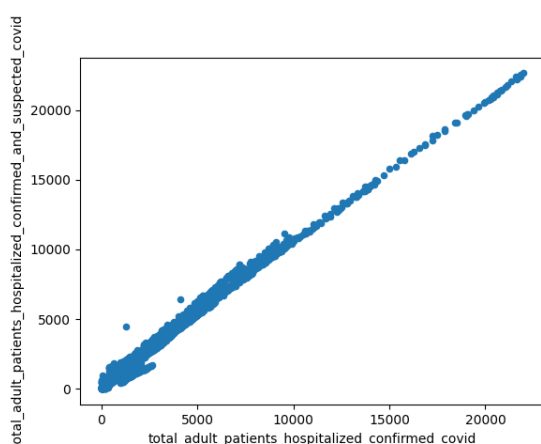
- critical_staffing_shortage_anticipated_within_week_yes (y)
- total_adult_patients_hospitalized_confirmed_covid (x)



The next graph presents the relationship between critical staff shortages anticipated within the week and the total number of adults that are confirmed to be hospitalized from Covid-19. As the line moves upward from left to right, we can see there is a strong positive relationship between the amount of shortages that hospitals face when the amount of Covid-19 patients they receive is increased.

As the spread of Covid-19 caused more and more patients to be sent to the hospital, the natural requirement of more able bodied nurses is directly seen here. However, it is important to compare the amount of confirmed cases versus the number of confirmed and suspected. As we can see in the next scatterplot shown.

Figure 1



Variables:

- total_adult_patients_hospitalized_confirmed_and_suspected_covid (y)
- total_adult_patients_hospitalized_confirmed_covid (x)

This scatterplot shows the relationship between the total number of adults that are confirmed and suspected to be hospitalized from Covid-19, with respect to those that are only confirmed to be hospitalized.



This represents the reported number of patients currently hospitalized in an adult inpatient bed who have laboratory-confirmed or suspected COVID-19, including those in observation beds, to those who are absolutely laboratory-confirmed with COVID-19.

This data allows us to see the positive correlation between these two attributes and how the suspected cases impact the confirmed cases.

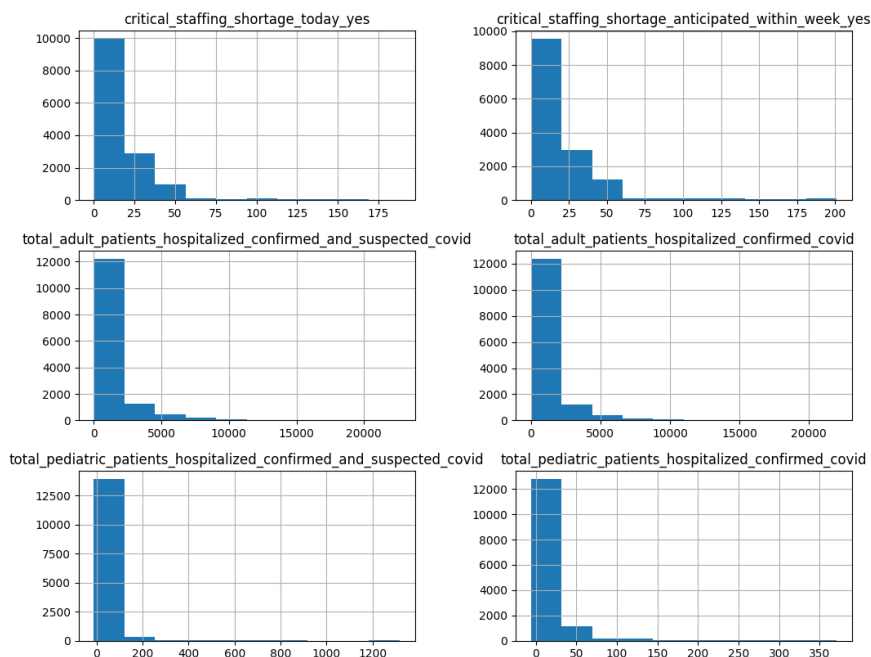
FEATURE SKEWNESS

Skewness (df.skew())

critical_staffing_shortage_today_yes	3.584149
critical_staffing_shortage_anticipated_within_week_yes	3.510674
total_pediatriac_patients_hospitalized_confirmed_and_suspected_covid_coverage	2.600400
total_pediatriac_patients_hospitalized_confirmed_covid	4.543498
total_pediatriac_patients_hospitalized_confirmed_covid_coverage	2.612634
total_staffed_adult_icu_beds_coverage	2.630415
inpatient_beds_utilization	-0.104708

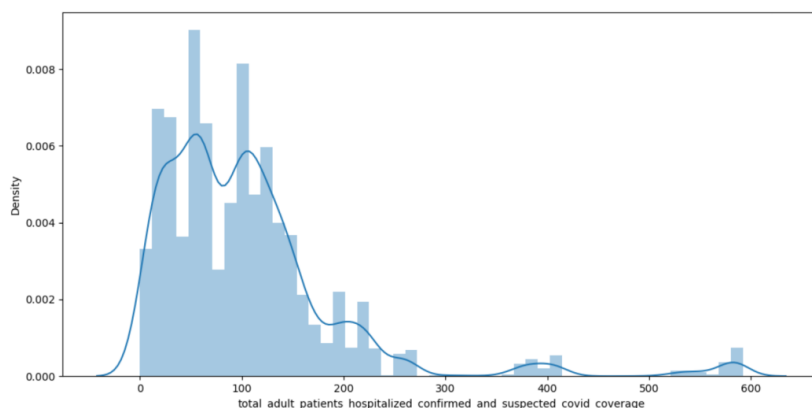
Asymmetrical and highly skewed, with the exception of patient bed utilization which is approximately symmetric

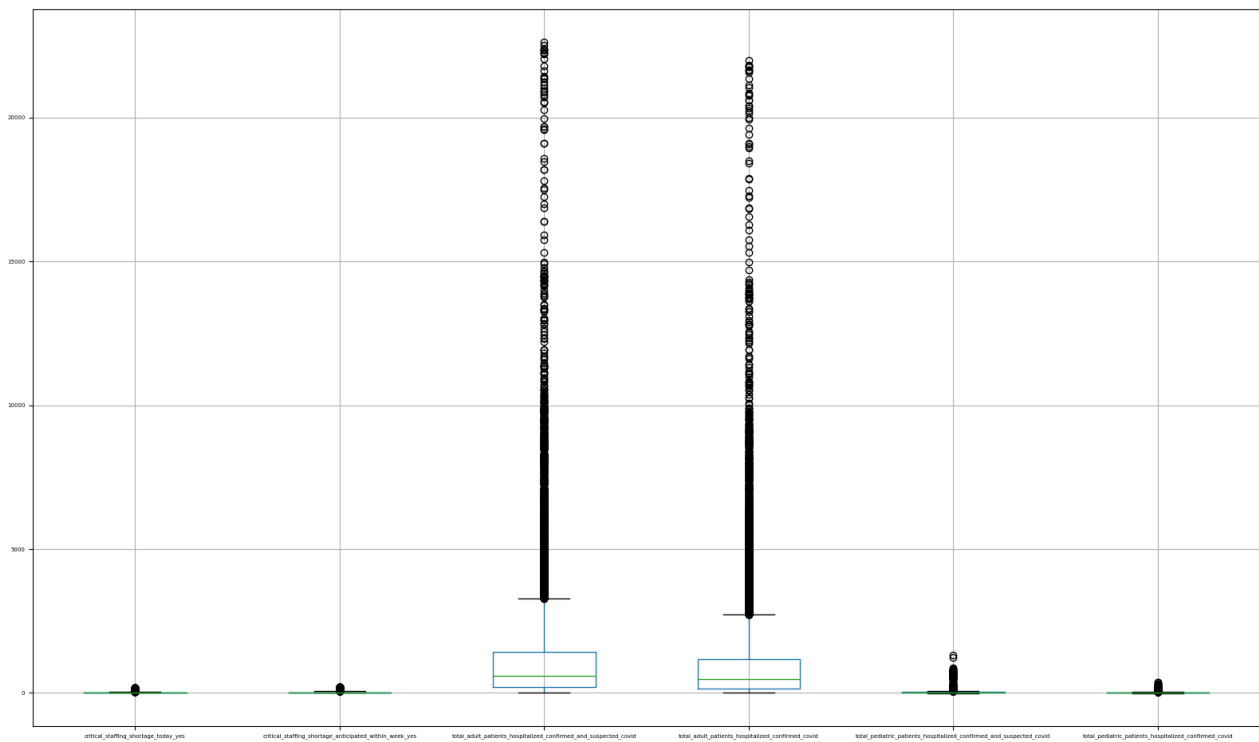
Another important observation is skewness of the features. We found that a majority of our data is highly skewed; out of 61 attributes, 60 is highly skewed and 1 is approximately symmetric. That column is inpatient bed utilization with a skewness of -0.104. Since the traits we want to look at are highly skewed, we are going to investigate their class distribution.



Inline with what we found with skewness, our data is positively skewed meaning a majority of the distribution is to the left, as shown in these histograms.

This univariate distribution plot shows the occurrence of the total number of hospitals where adult patients were hospitalized due to confirmed cases of COVID-19 or suspected cases. The kernel density estimation line shows the distribution of this attribute relative to the rest attributes.





T

Looking at the boxplot, we can see that information on staffing shortages, the two on the left, have outliers close to their medians similarly to the two attributes on the right, regarding the total number of pediatric patients confirmed and or suspected with covid.

The two distributions in the middle, however, have a majority of the outliers above the median line. This is in line with previous observations.

Conclusion

Information from this project has allowed us to identify key attributes that have a high correlation with staff shortages. This will allow us to create a prediction model based on these attributes.