MSDS 6372 Project 1
Data Set 1: Hospitalization Stays

Investigators:
Christian Castro, Victoria Hernandez, Troy McSimov

Introduction

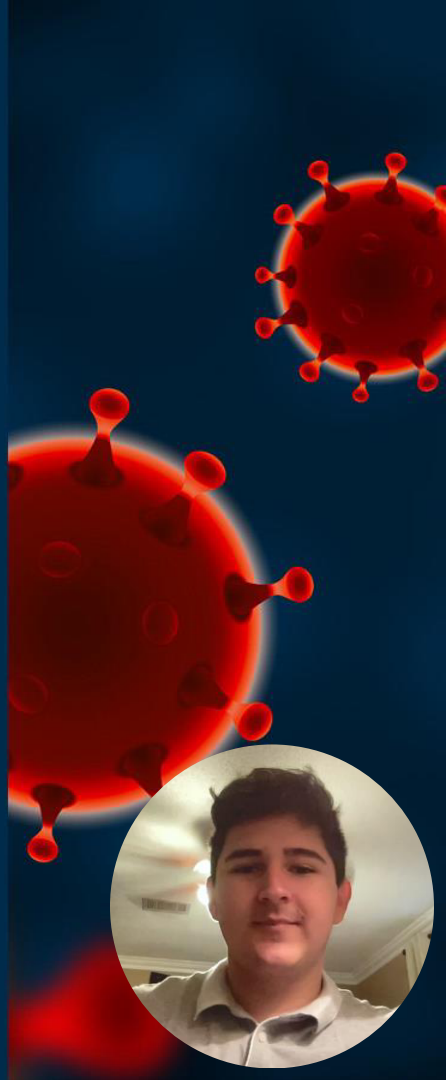
Meet the SMU Data Science team:

Christian Castro

Victoria Hernandez

Troy McSimo

Data Set 1: Hospitalization Stays

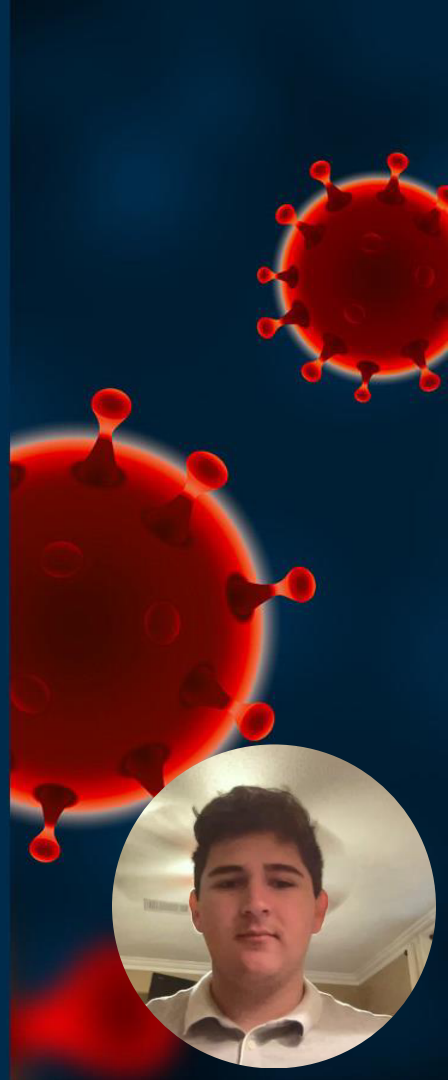


Objective Summary

Hospitals are constantly trying to understand and determine the factors that lead to long hospitalizations. Our analysis aims to identify these variables and predict their impact through an exploratory data analysis (EDA) and fulfill two objectives:

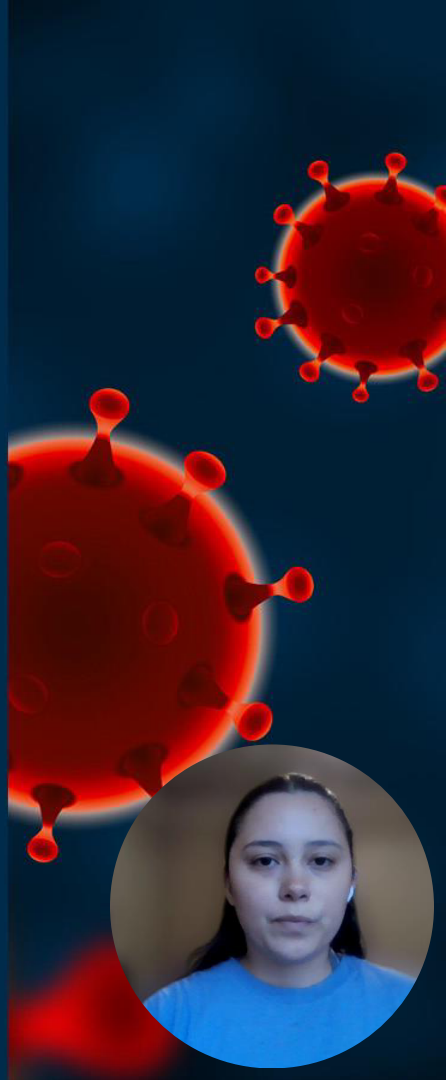
Objective 1: Identify correlations in the data that may contribute to the length of stay.

Objective 2: Create a predictive model to assist hospitals in forecasting the anticipated patient length of stay.



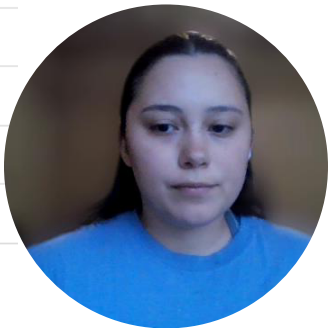
Data Overview

- The dataset "HospitalDurations.csv" has been provided
- 113 hospitals, 11 variables

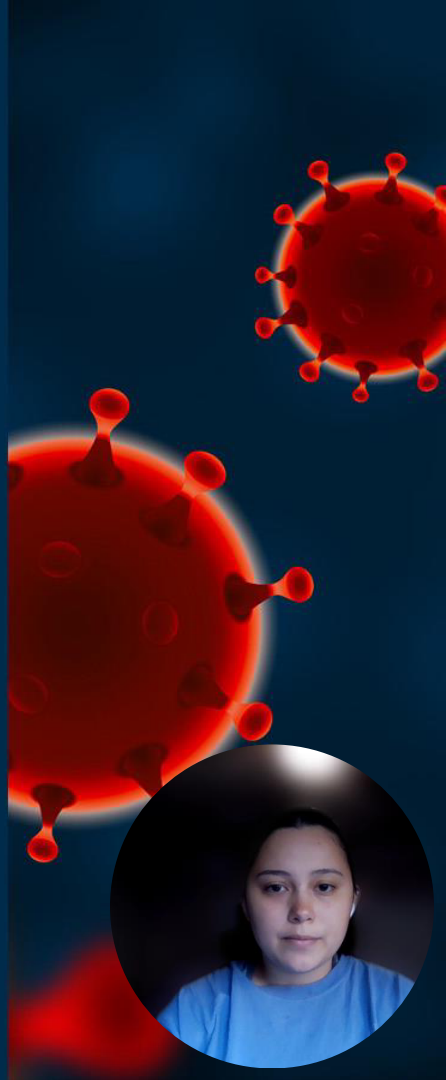


Data Description

Variable Name	Type	Description
ID	int	Record ID
Lgth.of.Sty	num	Average length of stay (in days)
Age	num	Average age of patients (in years)
Inf.Risk	num	Average estimated probability of hospital infection
R.Cul.Rat	num	Ratio of # of cultures taken to # of symptoms of infection x 100
R.CX.ray.Rat	num	Ratio of # of chest X-rays taken to # of symptoms of pneumonia x 100
N.Beds	int	Average number of beds
Med.Sc.Aff	int	Medical School Affiliation (1=Yes, 2=No)
Region	int	Region (1=NE, 2=NC, 3=S, 4=W)
Avg.Pat	int	Average number of patients in hospital per day
Avg.Nur	int	Average number of full time nurses
Pct.Ser.Fac	num	% of 35 potential facilities and services that are provided by the hospital



Exploratory Data Analysis



Data Tidying

Cleaning Data

- There were no variables with missing data fields or null values
- There were no duplicate rows

Data Types and Conversions

- Two variables were factored into categorical
- **Med.Sc.Aff** was categorized by ‘Affiliated’ or ‘Not-Affiliated’
- **Region** was categorized by ‘Northeast’, ‘Northcentral’ ‘South’, and ‘West’

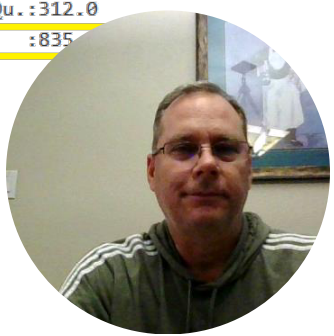
Data Summarization

- Length of Stay and Ratio of Chest X-ray appear to have outlier(s)
- Average Patients, Average Nurses, Ratio of Cultures Taken, and Number of Beds appear to be right skewed
- Disproportionate sample of patients in Affiliated vs Not-Affiliated hospitals

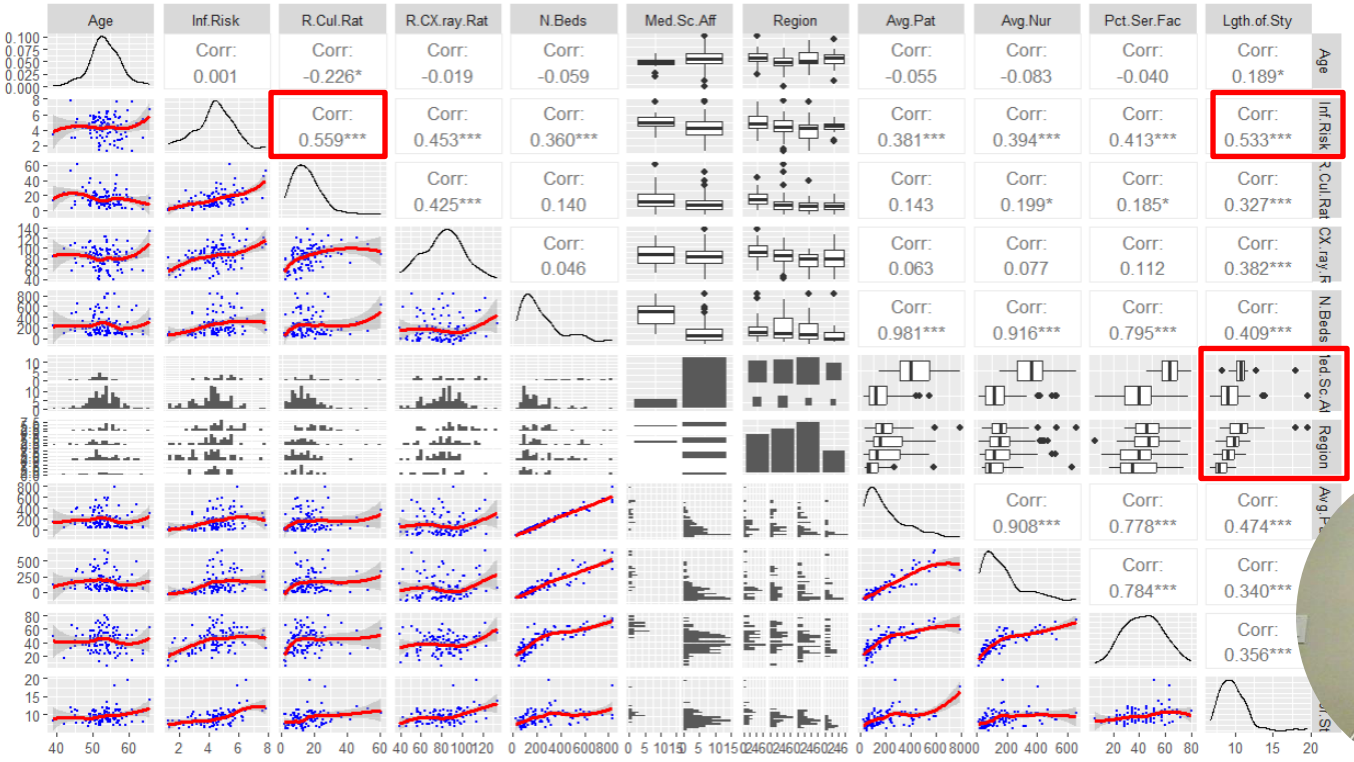
Lgth.of.Sty	Age	Inf.Risk	Pct.Ser.Fac	R.CX.ray.Rat
Min. : 6.700	Min. :38.80	Min. :1.300	Min. : 5.70	Min. : 39.60
1st Qu.: 8.340	1st Qu.:50.90	1st Qu.:3.700	1st Qu.:31.40	1st Qu.: 69.50
Median : 9.420	Median :53.20	Median :4.400	Median :42.90	Median : 82.30
Mean : 9.648	Mean :53.23	Mean :4.355	Mean :43.16	Mean : 81.63
3rd Qu.:10.470	3rd Qu.:56.20	3rd Qu.:5.200	3rd Qu.:54.30	3rd Qu.: 94.10
Max. :19.560	Max. :65.90	Max. :7.800	Max. :80.00	Max. :133.50

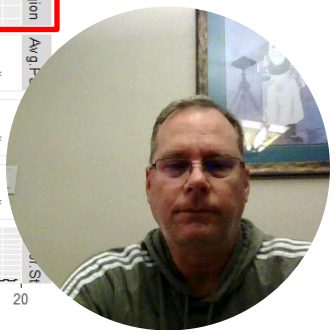
Avg.Pat	Avg.Nur	R.Cul.Rat	N.Beds
Min. : 20.0	Min. : 14.0	Min. : 1.60	Min. : 29.0
1st Qu.: 68.0	1st Qu.: 66.0	1st Qu.: 8.40	1st Qu.:106.0
Median :143.0	Median :132.0	Median :14.10	Median :186.0
Mean :191.4	Mean :173.2	Mean :15.79	Mean :252.2
3rd Qu.:252.0	3rd Qu.:218.0	3rd Qu.:20.30	3rd Qu.:312.0
Max. :791.0	Max. :656.0	Max. :60.50	Max. :835

Med.Sc.Aff	Region
Affiliated :17	Northeast :28
Not-Affiliated:96	Northcentral:32
	South :37
	West :16



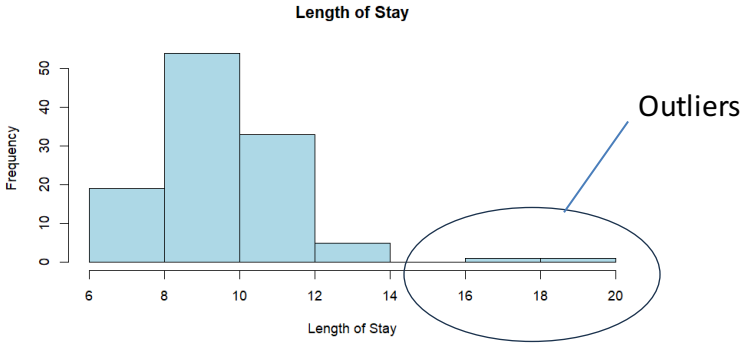
Data Analysis



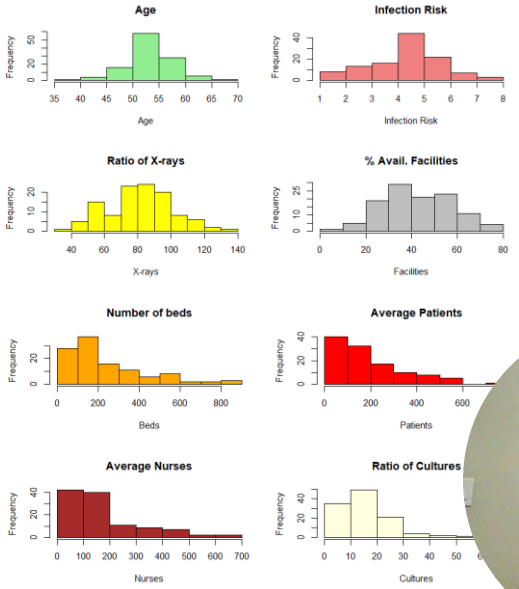


Visualizing with Histograms

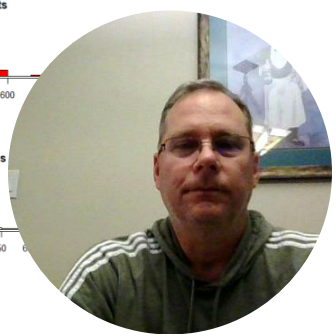
Dependent Variable



Predictor Variables

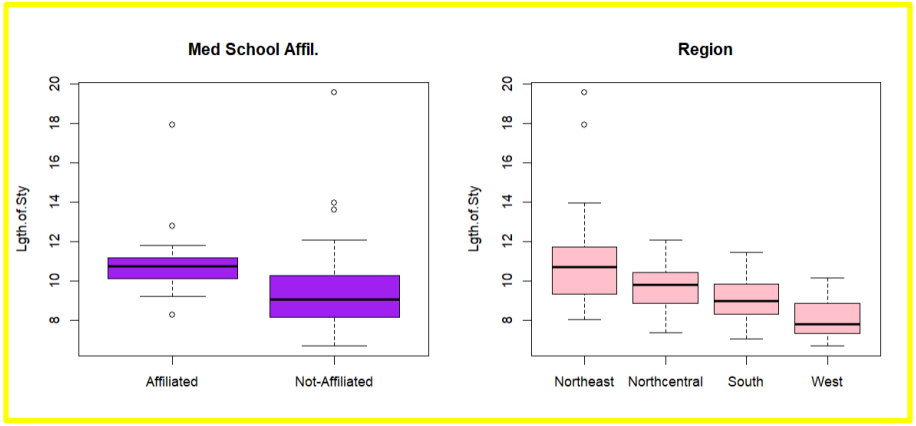


Right-skewed



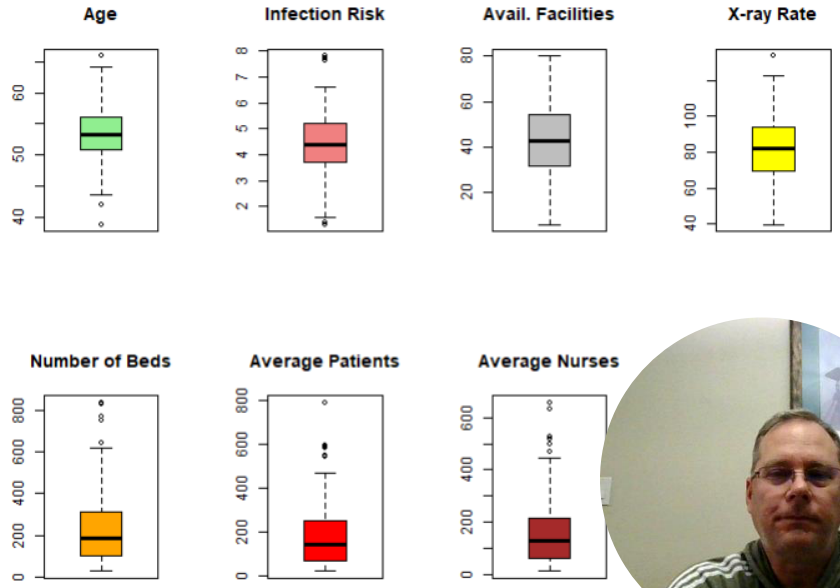
Visualizing Predictors with Box Plots

Categorical Data by Length of Stay



Both categorical predictors appear to have impact on length of stay

Numerical Data

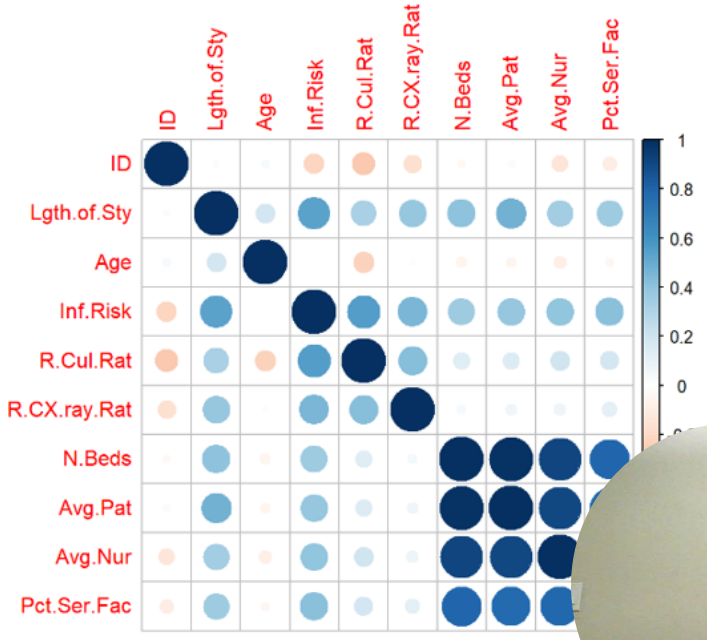


Checking for Collinearity

Infectious Risk had the highest correlation to Length of Stay, followed by Average # of Patients

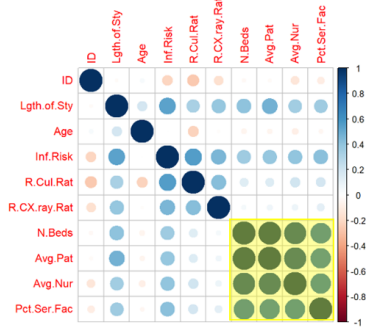
When looking at multicollinearity and what other variables may influence Infectious Risk, the **Ratio of Cultures** taken had the highest correlation, followed by the **Ratio of X-rays taken** and the percentage of **Services at the Facility**.

	ID	Lgth.of.Sty	Age	Inf.Risk	R.Cul.Rat	R.CX.ray.Rat	N.Beds	Avg.Pat	Avg.Nur	Pct.Ser.Fac
ID	1.0000000	-0.0223872	0.0372580	-0.2113157	-0.2674153	-0.1660204	-0.0356529	-0.0270560	-0.1335313	-0.0978575
Lgth.of.Sty	-0.0223872	1.0000000	0.1889140	0.5334438	0.3266838	0.3824819	0.4092652	0.4738855	0.3403671	0.3555379
Age	0.0372580	0.1889140	1.0000000	0.0010932	-0.2258468	-0.0188549	-0.0588232	-0.0547747	-0.0829446	-0.0404514
Inf.Risk	-0.2113157	0.5334438	0.0010932	1.0000000	0.5591589	0.4533916	0.3597700	0.3814111	0.3939813	0.4126007
R.Cul.Rat	-0.2674153	0.3266838	-0.2258468	0.5591589	1.0000000	0.4249620	0.1397249	0.1429482	0.1988998	0.1851311
R.CX.ray.Rat	-0.1660204	0.3824819	-0.0188549	0.4533916	0.4249620	1.0000000	0.0458200	0.0629135	0.0773813	0.1119276
N.Beds	-0.0356529	0.4092652	-0.0588232	0.3597700	0.1397249	0.0458200	1.0000000	0.9809977	0.9155042	0.7945244
Avg.Pat	-0.0270560	0.4738855	-0.0547747	0.3814111	0.1429482	0.0629135	0.9809977	1.0000000	0.9078970	0.7780633
Avg.Nur	-0.1335313	0.3403671	-0.0829446	0.3939813	0.1988998	0.0773813	0.9155042	0.9078970	1.0000000	0.7835055
Pct.Ser.Fac	-0.0978575	0.3555379	-0.0404514	0.4126007	0.1851311	0.1119276	0.7945244	0.7780633	0.7835055	1.0000000



Checking for Multicollinearity

There is significant correlation between the variables highlighted to the right although this may simply indicate that larger hospitals inherently have more beds, more nurses, more patients, and more services offered.



When looking at the VIF scores for evaluate the effects of multicollinearity, number of beds and average patients had the highest multicollinearity suggesting the possible benefit of leaving one of them out of the models.

Age: GVIF^(1/(2*Df)) = 1.08 (Low multicollinearity)

Inf.Risk: GVIF^(1/(2*Df)) = 1.47 (Low multicollinearity)

R.Cul.Rat: GVIF^(1/(2*Df)) = 1.41 (Low multicollinearity)

R.CX.ray.Rat: GVIF^(1/(2*Df)) = 1.19 (Low multicollinearity)

N.Beds: GVIF^(1/(2*Df)) = 5.97 (Moderate multicollinea

Med.Sc.Aff: GVIF^(1/(2*Df)) = 1.36 (Low multicollinea

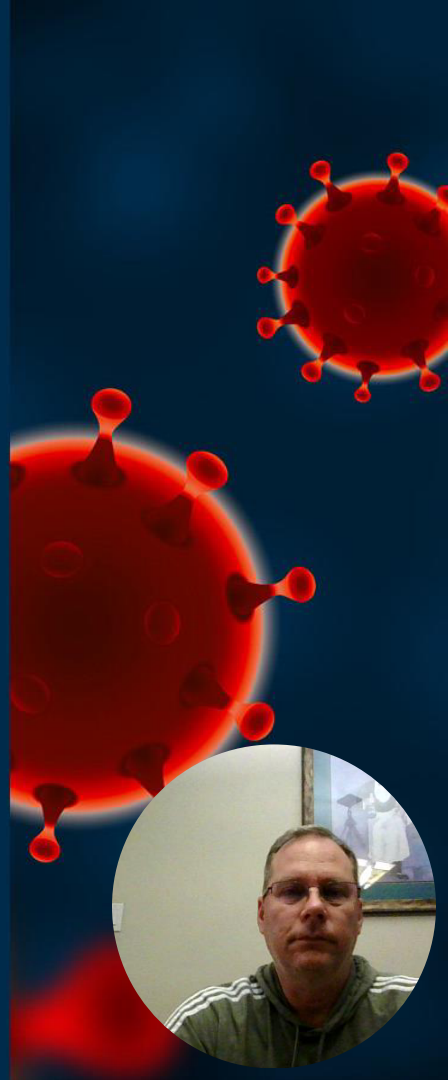
Region: GVIF^(1/(2*Df)) = 1.09 (Low multicollinearity

Avg.Pat: GVIF^(1/(2*Df)) = 5.85 (Moderate multicolli

Avg.Nur: GVIF^(1/(2*Df)) = 2.66 (Low multicollinearity)

Pct.Ser.Fac: GVIF^(1/(2*Df)) = 1.80 (Low multicollinearity)

Objective 1:
Is Infectious Risk Linked to Length of Hospital Stays?



Building the Model

- Our goal is to understand how the different variables influence the length of hospital stays with a particular focus on infection risk

Model: LengthOfStay ~ Age + InfectionRisk + RandomCulture + RandomXRay + NumberOfBeds + MedSchoolAffiliation + R
+ AveragePatients + AverageNurses + PercentFaciliti

The Use of LASSO Selection

Data Prep

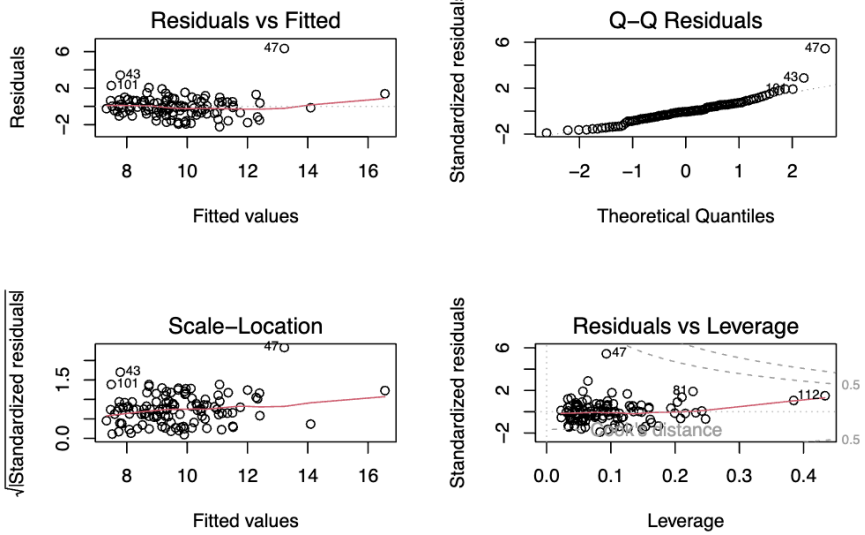
- We began by splitting the data 80/20 into training and testing sets respectively

- This allows us to use cross-validation to have a low lambda value (0.0025), preserving complexity in the model

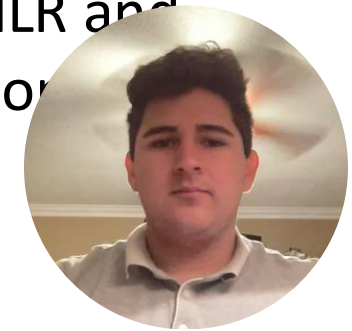
LASSO Results

- (Intercept): 3.7056
- ID: 0.0036
- Age: 0.0892
- Inf.Risk: 0.5175
- R.Cul.Rat: 0.0086
- R.CX.ray.Rat: 0.0070
- N.Beds: -0.0053
- Med.Sc.Aff: -0.395
- Region: -0.6047
- Avg.Pat: 0.0157
- Avg.Nur: -0.0065
- Pct.Ser.Fac: -0.006

Addressing the Assumptions



- The residual plot appears to be randomly distributed
- The QQ plot appears to be linear enough to satisfy the assumptions of MLR and LASSO selection



Model Performance

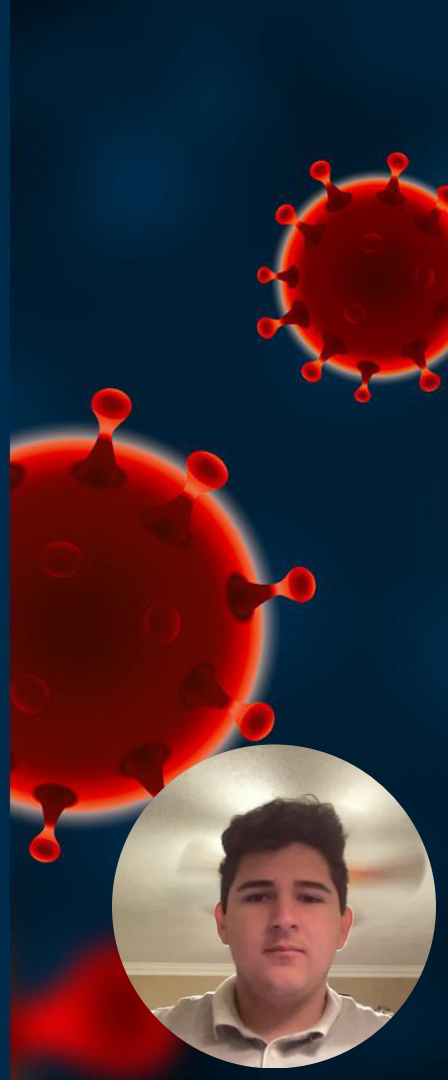
Test Set Results

- Mean Squared Error (MSE) 1.082
- Root Mean Squared Error (RMSE) 1.041

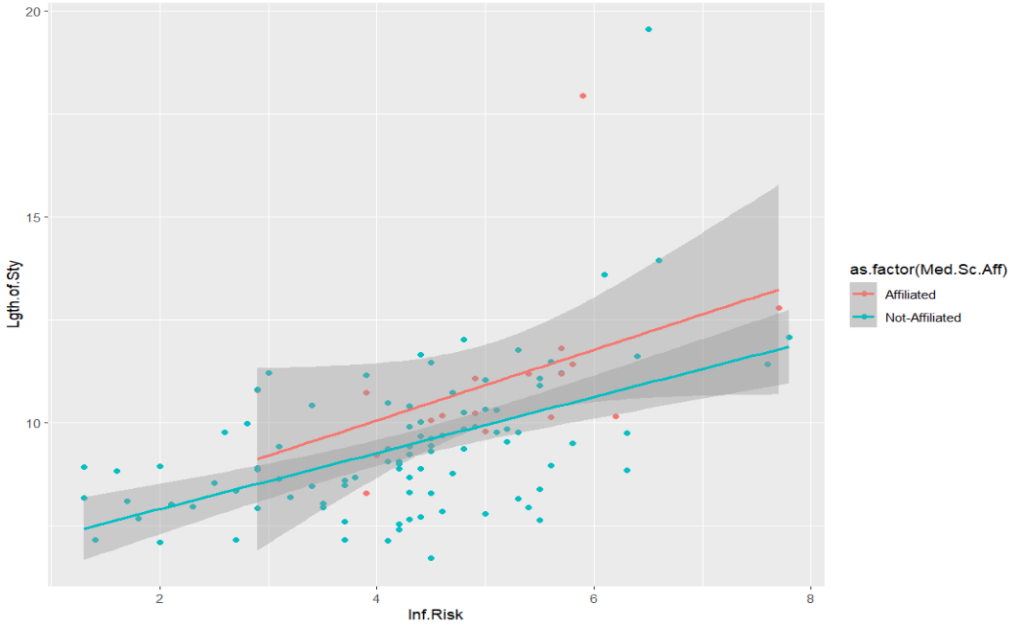
Analysis of Results

- The model explains a significant portion of the variance in the length of stay.
- Infection risk is positively associated with the length of stay (0.5175)

Objective 2:
Predictive Modeling to Forecast Anticipated Patient Length of Stay

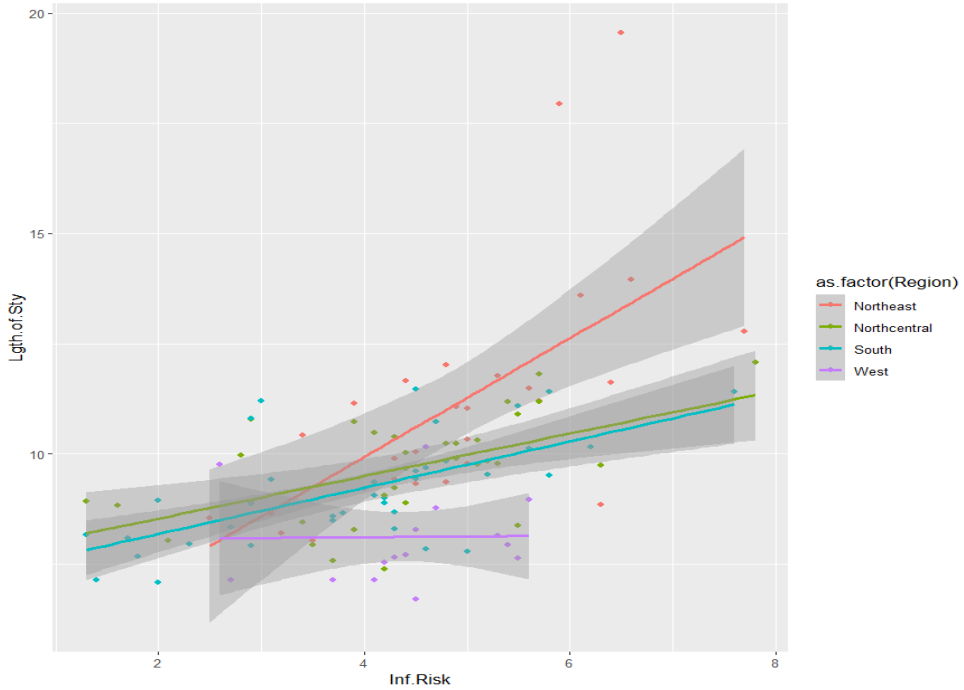


Visualize Potential Interactions

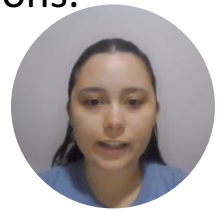


- Both affiliated and not affiliated show positive trend
- Slope for affiliated appears steeper than not- affiliated
- Confidence band for Affiliated hospitals seems broader, especially at higher infection risks, which could imply more variability in how infection risk impacts the length of stay at different affiliated hospitals.
- Considerable scatter around the trend lines, indicating variability that is not explained solely by infection risk

Visualize Potential Interactions



- Northeast, Northcentral, and South have positive linear slopes, whereas West is flat.
- Slope for Northeast appears steeper than the others
- Confidence band for Northeast region seems broader, especially at higher infection risks, which could imply more variability in how infection risk impacts the length of stay at different regions.



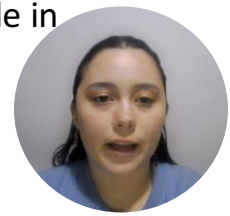
Complex MLR Model

Model: Lgth.of.Sty ~ Inf.Risk * Region + Avg.Pat + Avg.Nur + Age

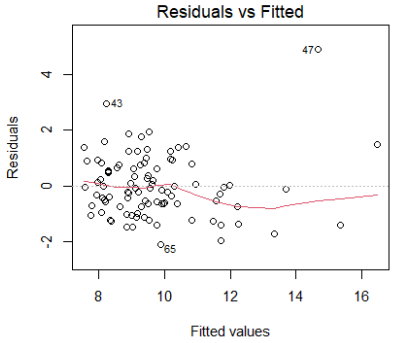
term	estimate	std.error	statistic	p.value
(Intercept)	1.2050838	1.7443137	0.6908642	0.4916020
Inf.Risk	1.4138975	0.2354521	6.0050326	0.0000001
RegionNorthcentral	3.2718503	1.3997768	2.3374086	0.0218563
RegionSouth	2.9286460	1.2857002	2.2778607	0.0253374
RegionWest	4.2298092	1.9855508	2.1302951	0.0361440
Avg.Pat	0.0084152	0.0020454	4.1141336	0.0000919
Avg.Nur	-0.0065308	0.0021540	-3.0318965	0.0032524
Age	0.0518733	0.0285447	1.8172641	0.0728297
Inf.Risk:RegionNorthcentral	-0.9731533	0.2950360	-3.2984222	0.0014395
Inf.Risk:RegionSouth	-0.9661044	0.2784601	-3.4694541	0.0008340
Inf.Risk:RegionWest	-1.4390503	0.4399574	-3.2708860	0.0015692

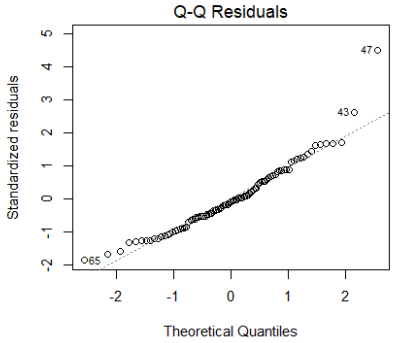
```
Residual standard error: 1.184 on 82 degrees of freedom
Multiple R-squared:  0.6909,    Adjusted R-squared:  0.6532
F-statistic: 18.33 on 10 and 82 DF,  p-value: < 2.2e-16
```

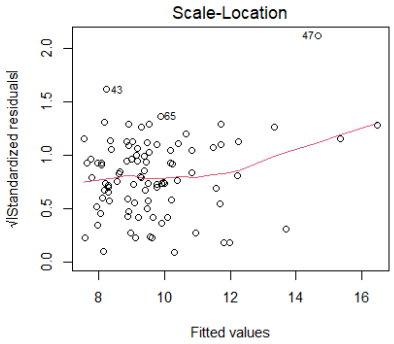
- `Avg.Nur`, `Avg.Pat`, and `Inf.Risk` were significant in determining hospital stays; with ‘Age’ showing lower significance.
- The model explains approximately 69% of the variance in hospital stay lengths, indicating a good fit.
- Notably, the interaction between infection risk and region suggests a potential relationship between these two predictors and their role in determining hospital stays.

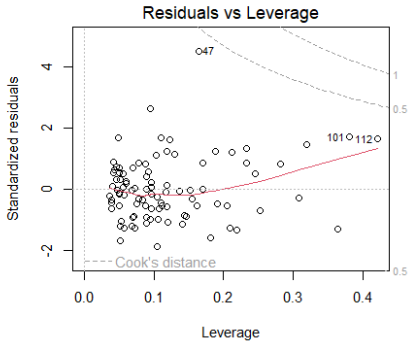


Complex MLR Model

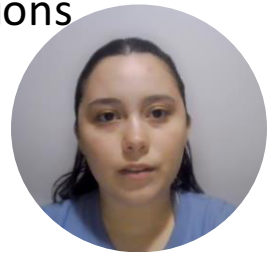




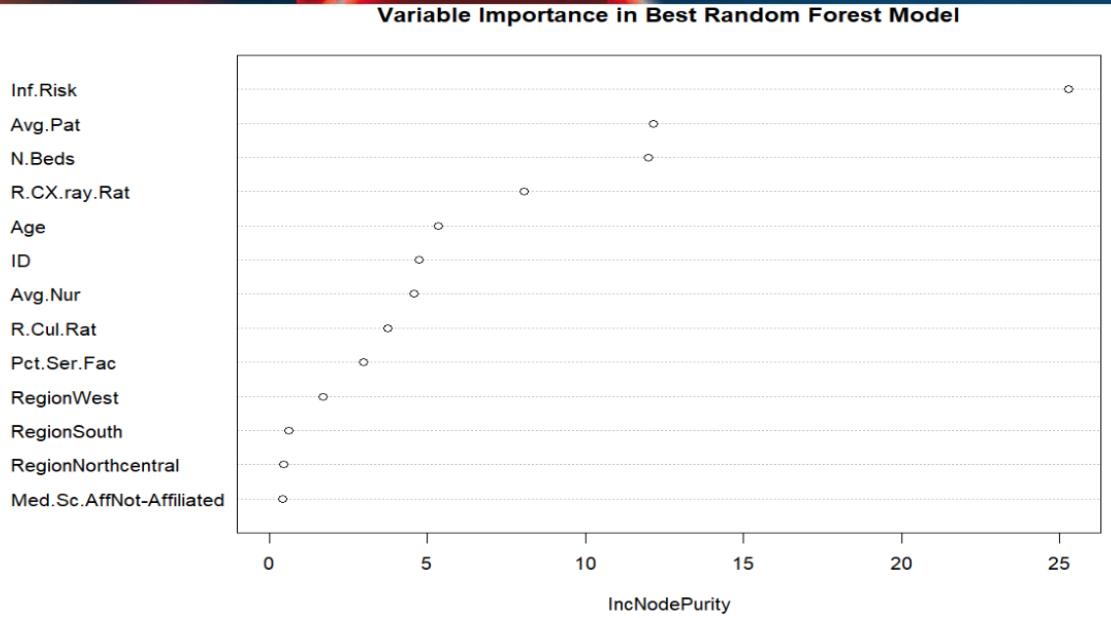




- The residual plot appears to be randomly distributed
- The QQ plot appears to be linear enough to satisfy the assumptions of MLR



Random Forest Model



Performance Summary of Best Tuned Random Forest Model

	Model	RMSE	R_Squared	Mean_Squared_Residuals	Percent_Variance_Explained
1	Best Tuned Random Forest	9.827195726281198	0.4987661620906966	1.152701903558131	49.87661620906966

- Most Influential variables appear to be 'Inf. Risk', 'Avg.Pat', and 'N.Beds'. This suggest these factors are critical in determining length of stay possibly due to direct impact on patient care and hospital resources.
- This model explains 49.8% of variance in the length of stay indicating a moderate fit.

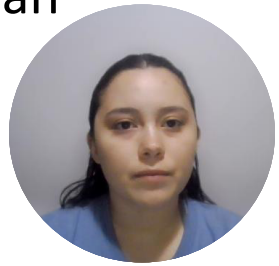
Model Performance Comparison

Model Performance Comparison

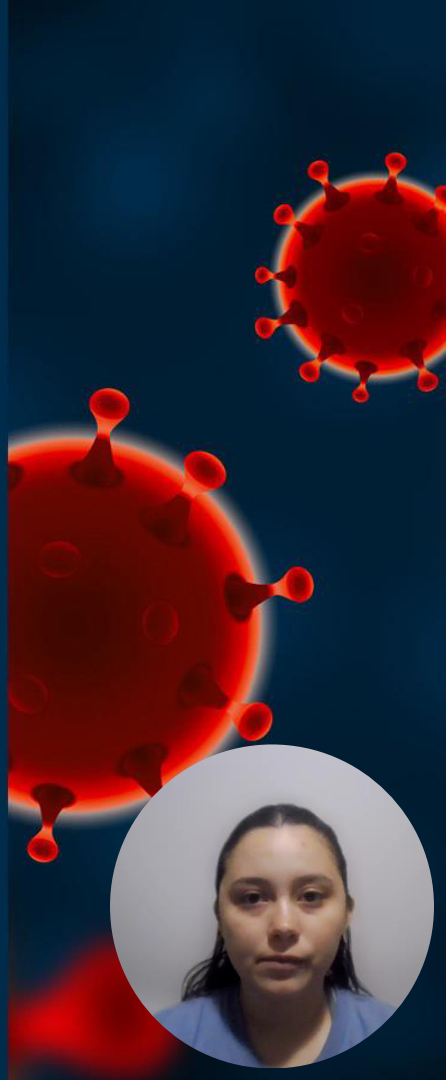
Model	RMSE
Complex MLR	4.083281
Random Forest	3.018700

The final comparison indicated the Random Forest is superior over the Complex MLR based on RMSE.

This suggests the Random Forest model may be capturing complex patterns more effectively than the MLR.



Conclusion



Objective 1 Conclusion

The Lasso regression model using cross-validation was used to select the best predictors for the model, which yielded an RMSE of 1.041.

- Infectious Risk
- Region
- Average Daily Number of Patients
- Average Number of Nurses
- Age

Objective 2 Conclusion

The data set was fit to two additional models and RMSE produced for comparison.

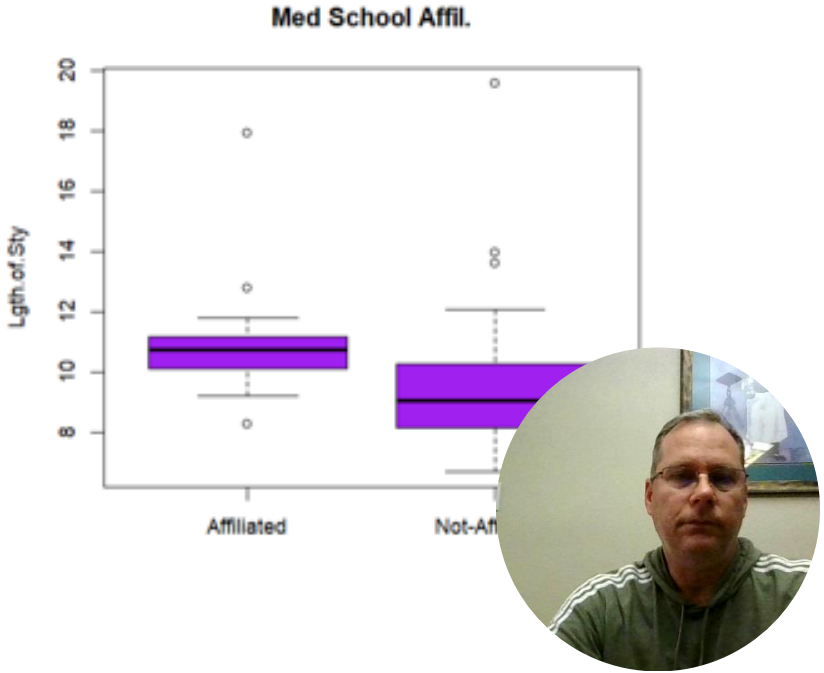
The Random Forest outperformed the MLR model, indicating a better fit model for predicting hospital stays with this data set.

Other Considerations

Additional Research Necessary

Medical School Affiliated hospitals are significantly larger (avg. 522 beds) than Non-Affiliated hospitals (avg. 204 beds) which equates to more patients, more nurses, and higher % of facility services, etc.

The box plot indicates that **Med.Sc.Aff** has a noticeable impact on Length of Stay or is this just noise introduced by multicollinearity as we saw between average # of beds and average # of patients?



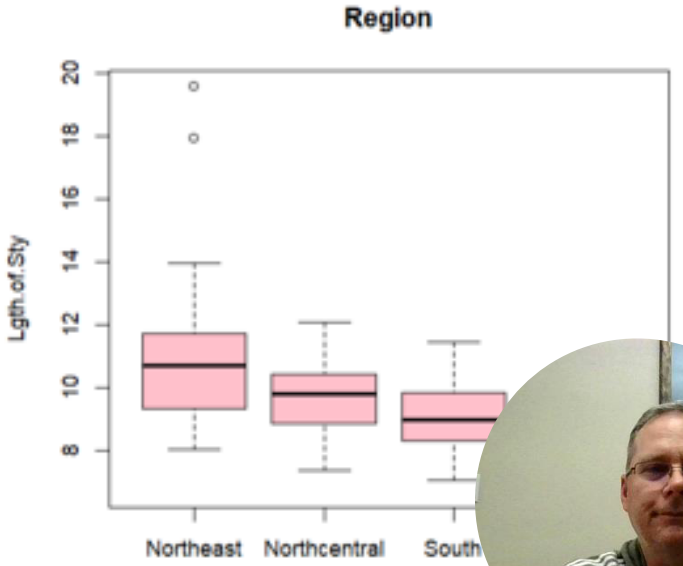
Other Considerations

Additional Research Necessary

What is the cause for Region having an impact on Length of Stay?

Could it be related to a virus outbreak more prevalent in particular regions?

Could it be related to other demographics like age, race, income?



Next Steps

Additional Modeling

With additional time, the team would like to continue running models to assist with predictor selection, transformations, and further tuning of our predicti model for increased accuracy.

Thank you

Questions and feedback can be sent via email to:

Christian Castro – ccastro@smu.edu

Victoria Hernandez – vhernandez@smu.edu

Troy McSimov – tmcsimov@smu.edu