Applied Data Science Portfolio Milestone

By Jose Reyes

Contents

Project Codes and Supporting Documents can be found on:

https://github.com/JoeR1221/Data-Science-Portfolio

Introduction

The Applied Data Science Program at Syracuse University offers a comprehensive set of courses in its curriculum that focuses on different areas of data science including data capture, management, analysis, and communication for decision-making. While these courses cover a variety topic from database management to machine learning, each one can be described and linked to one or more of the program's following learning goals:

1. Describe a broad overview of the major practice areas of data science.
2. Collect and organize data.
3. Identify patterns in data via visualization, statistical analysis, and data mining.
4. Develop alternative strategies based on the data.
5. Develop a plan of action to implement the business decisions derived from the analyses.
6. Demonstrate communication skills regarding data and its analysis for managers, IT professionals, programmers, statisticians, and other relevant professionals in their organization.
7. Synthesize the ethical dimensions of data science practice (e.g., privacy).

In this report, projects from several courses in the curriculum are described in detail to demonstrate how they collectively satisfy each of the program's overall learning goals. Additionally, the projects demonstrate how different applications of data science can extract information from increasingly large amounts of data that can ultimately lead to impactful decision making.

IST 707 - Data Analytics

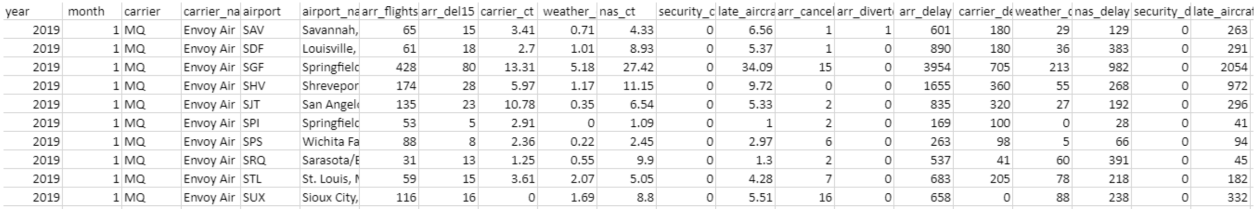
IST 707 focuses on data mining techniques and covers several classical machine learning algorithms including support-vector machine, k-means clustering, association rules mining, decision trees, and Naïve Bayes classifier. The final project for this course challenges students to apply these machine learning algorithms to an existing dataset and to describe the findings and draw conclusions from the results. In this project, 2017-2019 United States flight data was extracted from the Bureau of Transportation Statistics which include historical data of flights, carriers, and airports and their various delay and cancellation numbers.

Flight Data – Introduction

The goal of this analysis is to gain insights from airline travel data to better understand the chances of flight delays and cancellations. The likelihood of a delay or cancellation occurring can vary based on the arriving flights' city, carrier, and year.

It is estimated that flight delays cost $28.9 billion dollars due to lost time, opportunity costs, cancellations, and missed connections. Passengers bear more than half of the total cost. In 2019, passengers were delayed by more than 95 million minutes and approximately 20% of total flights in the United States were delayed by more than 15 minutes.

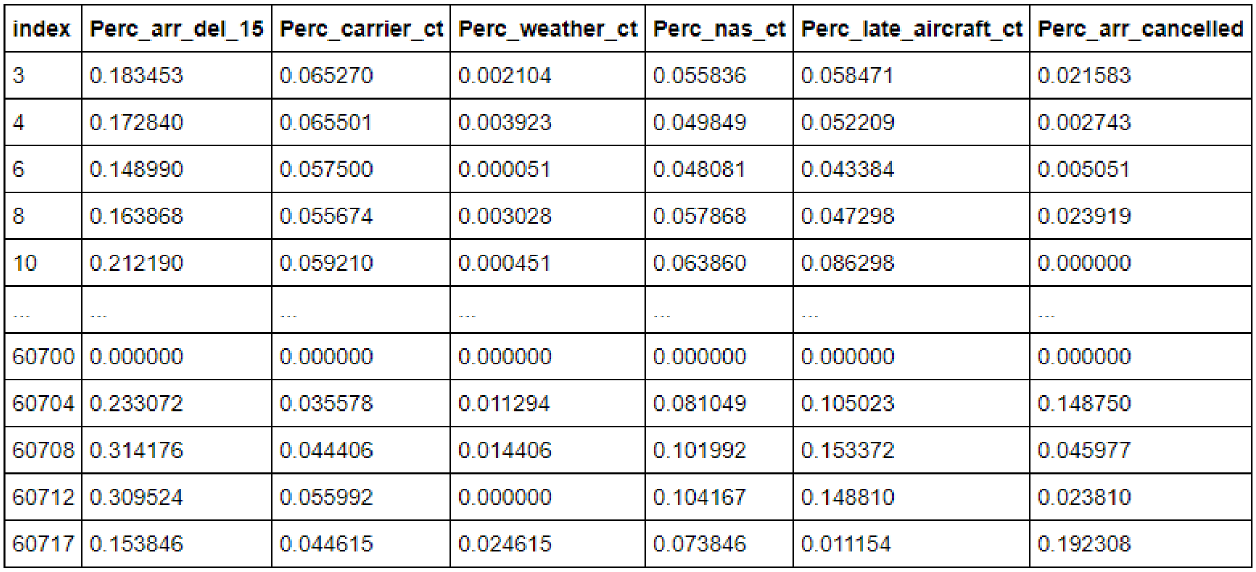
Before any data mining techniques can be conducted, the data must be collected and organized into a more usable form. While the raw flight data shown in Figure 1 is already organized into structured rows and columns, it is unusable in its current form for several data mining techniques.



year	month	carrier	carrier_na	airport	airport_na	arr_flights	arr_del15	carrier_ct	weather_	nas_ct	security_c	late_aircra	arr_cancel	arr_divert	arr_delay	carrier_de	weather_c	nas_delay	security_d	late_aircraf
2019	1	MQ	Envoy Air	SAV	Savannah,	65	15	3.41	0.71	4.33	0	6.56	1	1	601	180	29	129	0	263
2019	1	MQ	Envoy Air	SDF	Louisville,	61	18	2.7	1.01	8.93	0	5.37	1	0	890	180	36	383	0	291
2019	1	MQ	Envoy Air	SGF	Springfielc	428	80	13.31	5.18	27.42	0	34.09	15	0	3954	705	213	982	0	2054
2019	1	MQ	Envoy Air	SHV	Shrevepor	174	28	5.97	1.17	11.15	0	9.72	0	0	1655	360	55	268	0	972
2019	1	MQ	Envoy Air	SJT	San Angel	135	23	10.78	0.35	6.54	0	5.33	2	0	835	320	27	192	0	296
2019	1	MQ	Envoy Air	SPI	Springfielc	53	5	2.91	0	1.09	0	1	2	0	169	100	0	28	0	41
2019	1	MQ	Envoy Air	SPS	Wichita Fa	88	8	2.36	0.22	2.45	0	2.97	6	0	263	98	5	66	0	94
2019	1	MQ	Envoy Air	SRQ	Sarasota/E	31	13	1.25	0.55	9.9	0	1.3	2	0	537	41	60	391	0	45
2019	1	MQ	Envoy Air	STL	St. Louis, N	59	15	3.61	2.07	5.05	0	4.28	7	0	683	205	78	218	0	182
2019	1	MQ	Envoy Air	SUX	Sioux City,	116	16	0	1.69	8.8	0	5.51	16	0	658	0	88	238	0	332

Figure 1. Sample of raw flight data from the United States Bureau of Transportation Statistics

The motivation in this project is to identify patterns in flight delays and cancellation among the different airports, airline carriers, and time of year. To run this data through an algorithm such as k-means clustering and association rules mining (ARM), several data organizing and manipulation steps need to be performed.

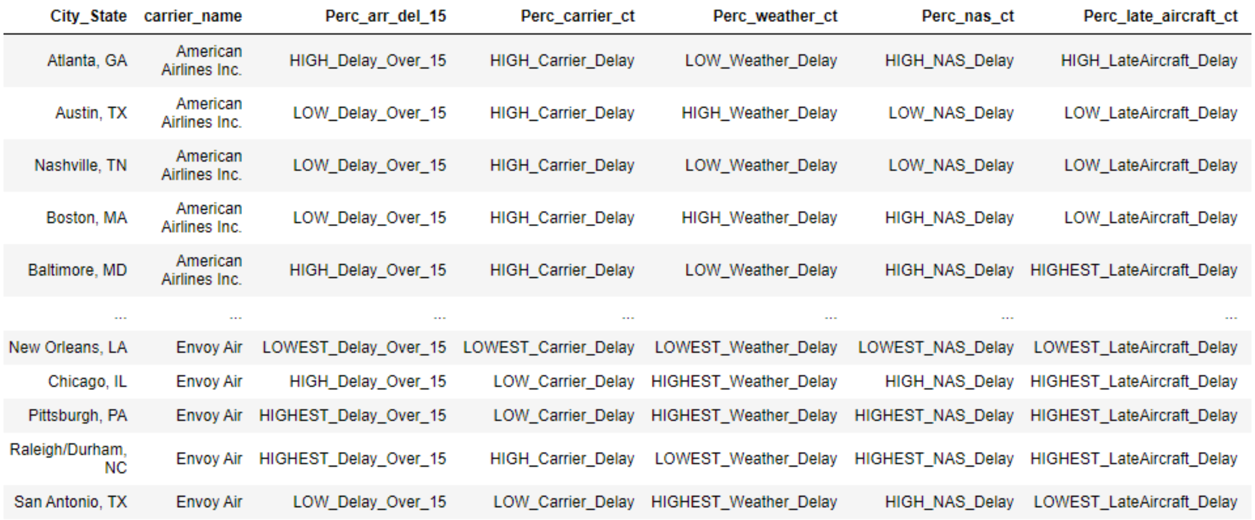


index	Perc_arr_del_15	Perc_carrier_ct	Perc_weather_ct	Perc_nas_ct	Perc_late_aircraft_ct	Perc_arr_cancelled
3	0.183453	0.065270	0.002104	0.055836	0.058471	0.021583
4	0.172840	0.065501	0.003923	0.049849	0.052209	0.002743
6	0.148990	0.057500	0.000051	0.048081	0.043384	0.005051
8	0.163868	0.055674	0.003028	0.057868	0.047298	0.023919
10	0.212190	0.059210	0.000451	0.063860	0.086298	0.000000
...	...	...	...	...	...	...
60700	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
60704	0.233072	0.035578	0.011294	0.081049	0.105023	0.148750
60708	0.314176	0.044406	0.014406	0.101992	0.153372	0.045977
60712	0.309524	0.055992	0.000000	0.104167	0.148810	0.023810
60717	0.153846	0.044615	0.024615	0.073846	0.011154	0.192308

Figure 2. Sample of flight data after data cleansing for k-means clustering

The data shown in Figure 2 has been cleansed and can be used for k-means clustering on Python using the Pandas package. Because the raw data included flight data from 2017-2020, the averages were calculated for the percentage of airline delays that are more than 15 minutes (Perc_arr_del_15), delays due to carrier (Perc_carrier_ct), delays due to weather (Perc_weather_ct), delays due to national security advisories (Perc_nas_ct), delays due to aircraft (Perc_late_aircraft_ct), and percentage of flight cancellations (Perc_arr_cancelled) for each airport, airline carrier, and time of year. Additionally, each column was normalized so that each feature has equal weight in the k-means clustering.

Another way to organize this data is through the creation of discretized bins. This allows the data to be used for ARM which is an algorithm that produces a list of rules that each have a set of items that commonly occur together. Figure 3 shows the data after the feature values have been placed into discretized bins based on their relative values.



City_State	carrier_name	Perc_arr_del_15	Perc_carrier_ct	Perc_weather_ct	Perc_nas_ct	Perc_late_aircraft_ct
Atlanta, GA	American Airlines Inc.	HIGH_Delay_Over_15	HIGH_Carrier_Delay	LOW_Weather_Delay	HIGH_NAS_Delay	HIGH_LateAircraft_Delay
Austin, TX	American Airlines Inc.	LOW_Delay_Over_15	HIGH_Carrier_Delay	HIGH_Weather_Delay	LOW_NAS_Delay	LOW_LateAircraft_Delay
Nashville, TN	American Airlines Inc.	LOW_Delay_Over_15	HIGH_Carrier_Delay	LOW_Weather_Delay	LOW_NAS_Delay	LOW_LateAircraft_Delay
Boston, MA	American Airlines Inc.	LOW_Delay_Over_15	HIGH_Carrier_Delay	HIGH_Weather_Delay	HIGH_NAS_Delay	LOW_LateAircraft_Delay
Baltimore, MD	American Airlines Inc.	HIGH_Delay_Over_15	HIGH_Carrier_Delay	LOW_Weather_Delay	HIGH_NAS_Delay	HIGHEST_LateAircraft_Delay
...	...	...	...	...	...	...
New Orleans, LA	Envoy Air	LOWEST_Delay_Over_15	LOWEST_Carrier_Delay	LOWEST_Weather_Delay	LOWEST_NAS_Delay	LOWEST_LateAircraft_Delay
Chicago, IL	Envoy Air	HIGH_Delay_Over_15	LOW_Carrier_Delay	HIGHEST_Weather_Delay	HIGH_NAS_Delay	HIGHEST_LateAircraft_Delay
Pittsburgh, PA	Envoy Air	HIGHEST_Delay_Over_15	LOW_Carrier_Delay	HIGHEST_Weather_Delay	HIGHEST_NAS_Delay	HIGHEST_LateAircraft_Delay
Raleigh/Durham, NC	Envoy Air	HIGHEST_Delay_Over_15	HIGH_Carrier_Delay	LOWEST_Weather_Delay	HIGHEST_NAS_Delay	HIGHEST_LateAircraft_Delay
San Antonio, TX	Envoy Air	LOW_Delay_Over_15	LOW_Carrier_Delay	HIGHEST_Weather_Delay	HIGH_NAS_Delay	LOWEST_LateAircraft_Delay

Figure 3. Sample of flight data after data cleansing for association rules mining

Flight Data - K-means clustering results

There are several ways to visualize the k-means clustering results. The main goal from this technique is to find groups in the data to find airports, airline carriers, and time of year where flight delays and cancellations are most or least likely to occur.

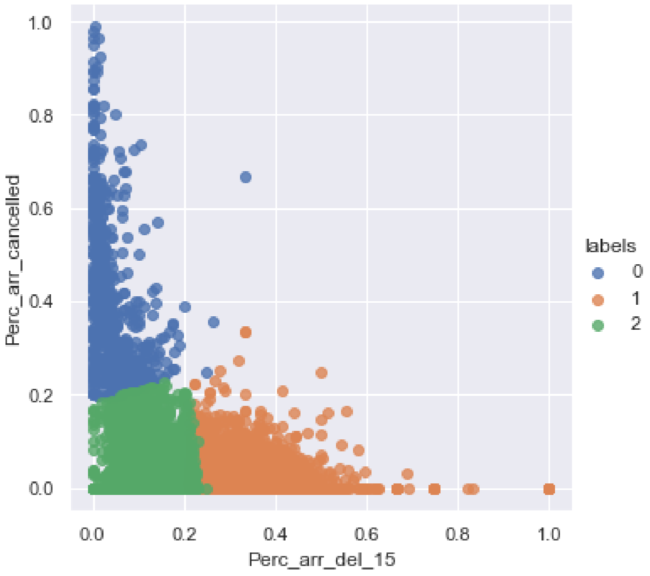


Figure 4. Two-dimensional scatterplot of K-means clustering results

Figure 4 displays a scatterplot of the results for two dimensions: Percentage of flight cancellations and flights that are delayed for over 15 minutes. Flights in cluster 0 have higher cancellation and lower delay percentages. Examples in this cluster include flying in Miami in April or New Orleans in December. Flights in Cluster 1 have higher delays and lower cancellation percentages which includes flights to Seattle in November or Boston in January. Finally, flights in cluster 2 include low cancellation and delay percentages. This includes flying to Baltimore in April or New York in May.

Figure 5 offers yet another way of visualizing the k-means clustering results; each scatterplot separates the clustering results for each specific feature.

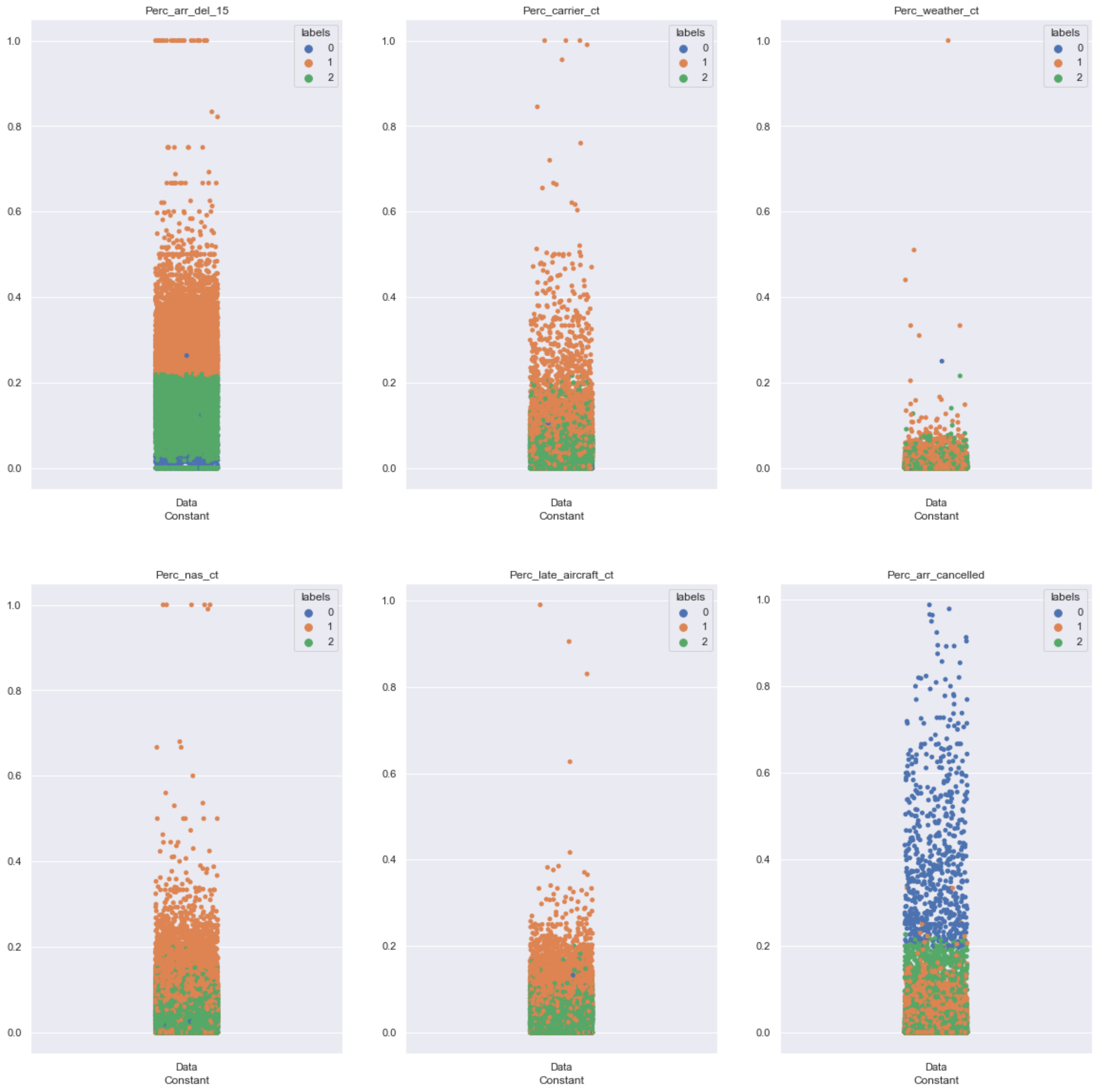
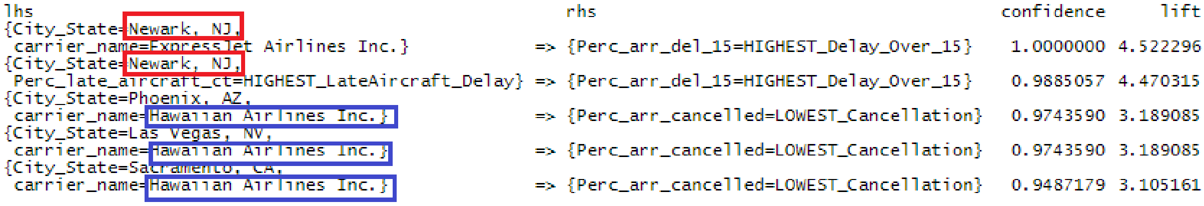


Figure 5. K-means clustering results of the flight data for two different dimensions

Flight Data - Association rules mining results

The ARM analysis focuses on rules with high confidence and lift scores, which respectively measure how frequently certain items appear with other items in an association and the ratio of the confidence of the rule and the expected confidence of the rule. Due to the large variety of airports, carriers, and features, the support for most rules is low.

Figure 6 shows the results when mining for rules with the highest percentages of delays over 15 minutes (bin name: HIGHEST_Delay_Over_15) or most elevated rate of cancellations (bin name: Perc_arr_cancelled=LOWEST_Cancellation) based on lift and confidence values.



```
lhs                                                  rhs                                        confidence     lift
{City_State=Newark, NJ,
 carrier_name=ExpressJet Airlines Inc.}           => {Perc_arr_del_15=HIGHEST_Delay_Over_15}    1.0000000 4.522296
{City_State=Newark, NJ,
 Perc_late_aircraft_ct=HIGHEST_LateAircraft_Delay} => {Perc_arr_del_15=HIGHEST_Delay_Over_15}    0.9885057 4.470315
{City_State=Phoenix, AZ,
 carrier_name=Hawaiian Airlines Inc.}             => {Perc_arr_cancelled=LOWEST_Cancellation}  0.9743590 3.189085
{City_State=Las Vegas, NV,
 carrier_name=Hawaiian Airlines Inc.}             => {Perc_arr_cancelled=LOWEST_Cancellation}  0.9743590 3.189085
{City_State=Sacramento, CA,
 carrier_name=Hawaiian Airlines Inc.}             => {Perc_arr_cancelled=LOWEST_Cancellation}  0.9487179 3.105161
```

Figure 6: ARM results for highest delays and cancellations

The top rule includes flights arriving in Newark, NJ, Express Jet Airlines, and having the highest delays. This rule also has the maximum confidence score meaning that 100% of Express Jet Airlines arriving in Newark have some of the highest percentages of flight delays.

Figure 7 shows these results created in R as a matrix where the size of the plot is confidence and color is lift. The top rules associated with Newark, NJ, and high delays indicate the highest confidence and lift values. When limiting the right-hand side (RHS) to the highest percentages of carrier-related delays, Hawaiian Airlines makes an appearance in all 5 of the top rules when sorted by confidence and lift. The flights that arrived in San Jose, CA, and Seattle, WA, all produced maximum confidence scores.

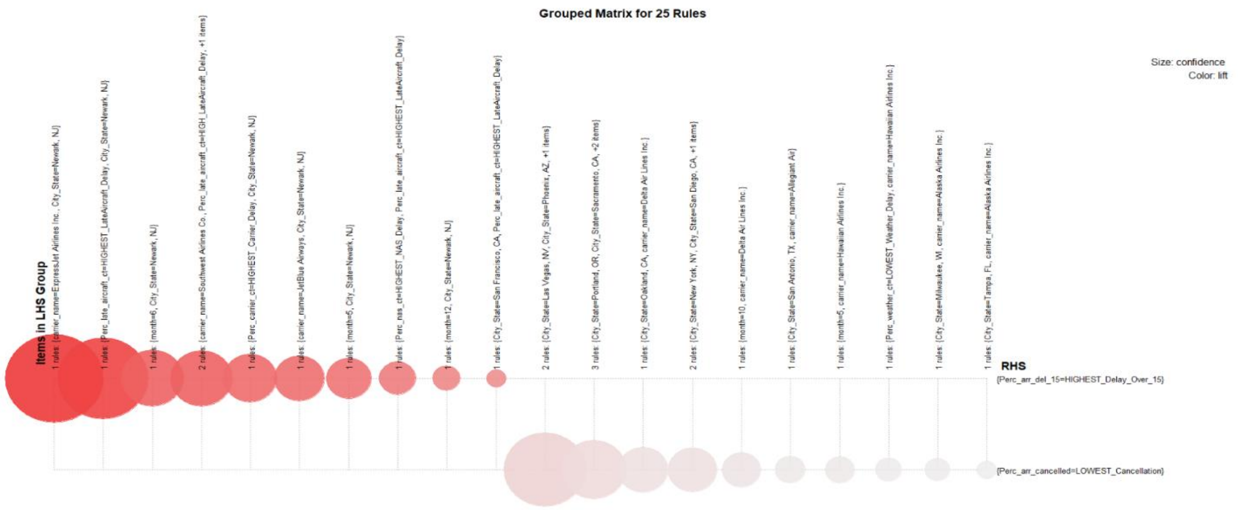


Figure 7: Graphed ARM results for highest delays and cancellations

Perhaps the most important aspect of these results is their ability to identify problem areas in the flight industry. In the clustering analysis, flights in March and April are overwhelmingly overrepresented in Cluster 0 where flights generally have high cancellations while flights that occur in the summer months are overwhelmingly overrepresented in Cluster 1. This can lead to a targeted approach when identifying improvement opportunities to airline efficiency - What is happening during these months that are causing cancellations and delays? Can it be linked to problems due to the aircraft, airport operations, or weather?

Conversely, it can identify what is going well in the industry. Flights in November are overwhelmingly represented in Cluster 2 where flights have relatively low delays and cancellations. Is this due to lower airport traffic? How can flights in other months be more like flights in November?

These findings can be useful for consumers as well. Clearly, flights in Newark appear to have issues based on the ARM results where flights through certain airline carriers are almost certainly going to be delayed. In essence, it would be prudent to avoid flights through this carrier in Newark if possible.

Flight Data - Linkage to Program Goals

Describe a broad overview of the major practice areas of data science – One of the key aspects of data science is its ability to gather insights that can ultimately lead to impactful decision making. This is demonstrated by the machine learning techniques shown in this project which used data mining techniques to extract interesting and previously unknowledge through a non-trivial task. This is what ultimately separates data mining tasks and machine learning algorithms from more trivial tasks such as a simple filter or query. Another key aspect of data science is its applicability to large data sets. The flight data contains 10,000 rows with 42 features and trivial tasks are not nearly as effective at finding hidden information when compared to non-trivial data mining techniques. This ability to extract hidden insights while, at the same time, managing large quantities of data are what makes data science applications so powerful and impactful in today's world.

Collect and organize data – Data is often not usable in its raw form for data analysis. Data cleansing can be an excruciatingly long process, but it is just as important as any other step in the data analysis. While the flight data was initially already structured, it still required several data cleansing steps including filtering, merging, normalizing, and discretizing. The data collecting and organizing demonstrated in this project shows how the data collection and organization process is conducted and that none of the following analysis could have been possible without it.

Develop alternative strategies based on the data - The analysis provided key results that can ultimately lead to the development of an action plan to improve efficiency and operations in the airline industry. As discussed earlier, the flight data results can identify problem areas associated with airline delays and cancellations. What airports or airline carriers are having issues? Is it the time of the year? Is it a combination of all these factors? This identification of specific problem areas can lead to a strategy for the airline industry and create improvement opportunities. It can also create strategies for travellers for the next time they plan a trip.

IST 718 - Big Data Analytics

In addition to classical machine learning algorithms, IST 718 introduces newer data science techniques such as neural networks and time series forecasting including autoregressive integrated moving average (ARIMA) and Prophet. The class project challenges students to use these advanced data science techniques to create a model that can lead to business solutions and decision making. In this example, a time series forecasting model was created to predict Covid-19 related deaths

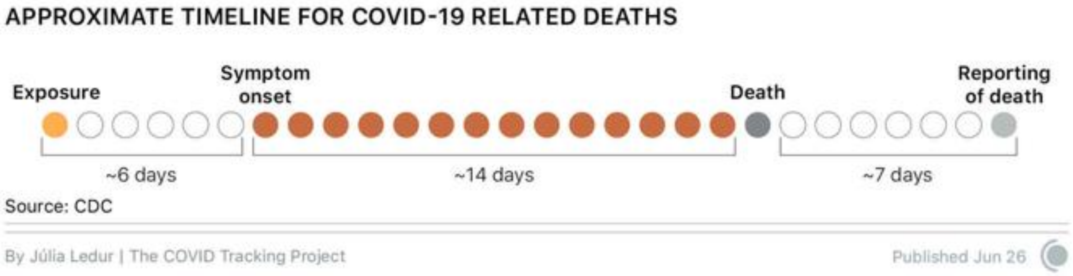
Covid-19 Deaths – Introduction

The goal of this project is to create a time series model that predicts Covid-19 related deaths using Prophet which is an additive regression model created by Facebook. It is particularly strong at forecasting data with daily observations and strong seasonalities. This project was completed in the summer of 2020 when the world was a little more uncertain of the pandemic's trajectory in the coming months.

Covid-19 Deaths - Time series forecasting using Prophet

The data used in this model comes from the COVID Tracking Project from The Atlantic. This dataset refreshes daily and includes each state's daily Covid-19 cases, deaths, hospitalization, and several other variables.

Prophet can create forecasting predictions based on historic trends as well as additional regressor values which can be added as part of the model. As a predictor of Covid-19 deaths, the most obvious regressor value is Covid-19 infections which generally precedes Covid-19 deaths by about 14 days. This is displayed by the timeline in Figure 8 from exposure to the onset of symptoms to the report of deaths. The time difference between the onset of symptoms and the reporting of a Covid-19 related death is about 21 days.



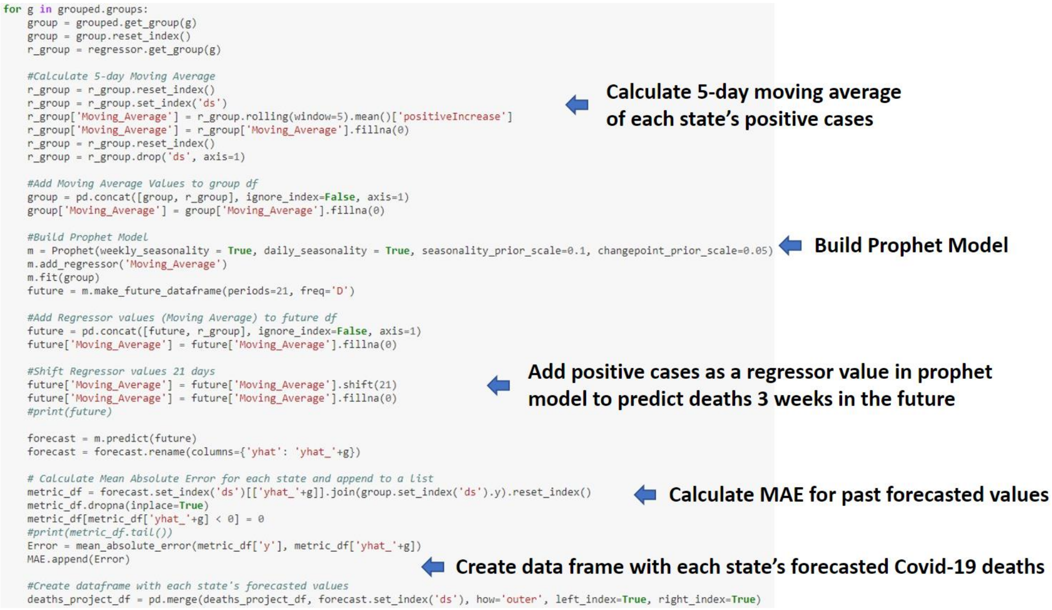
Source: CDC

By Júlia Ledur | The COVID Tracking Project

Published Jun 26

Figure 8: Timeline of Covid-19 related deaths from exposure to the reporting of death

To account for this variable range in lag between cases and deaths, the Prophet Model includes a 5-day moving average of reported positive coronavirus cases as regressor values. This value is then shifted 21 days to compare deaths with the reported cases 3 weeks prior. Figure 9 shows the components used in the building of a Prophet model in Python to predict Covid-19 deaths. The forecasting results predict the Covid-19 deaths for each state 21 days into the future.



```
for g in grouped.groups:
    group = grouped.get_group(g)
    group = group.reset_index()
    r_group = regressor.get_group(g)

    #Calculate 5-day Moving Average
    r_group = r_group.reset_index()
    r_group = r_group.set_index('ds')
    r_group['Moving_Average'] = r_group.rolling(window=5).mean()['positiveIncrease']
    r_group['Moving_Average'] = r_group['Moving_Average'].fillna(0)
    r_group = r_group.reset_index()
    r_group = r_group.drop('ds', axis=1)

    #Add Moving Average Values to group df
    group = pd.concat([group, r_group], ignore_index=False, axis=1)
    group['Moving_Average'] = group['Moving_Average'].fillna(0)

    #Build Prophet Model
    m = Prophet(weekly_seasonality = True, daily_seasonality = True, seasonality_prior_scale=0.1, changepoint_prior_scale=0.05)
    m.add_regressor('Moving_Average')
    m.fit(group)
    future = m.make_future_dataframe(periods=21, freq='D')

    #Add Regressor values (Moving Average) to future df
    future = pd.concat([future, r_group], ignore_index=False, axis=1)
    future['Moving_Average'] = future['Moving_Average'].fillna(0)

    #Shift Regressor values 21 days
    future['Moving_Average'] = future['Moving_Average'].shift(21)
    future['Moving_Average'] = future['Moving_Average'].fillna(0)
    #print(future)

    forecast = m.predict(future)
    forecast = forecast.rename(columns={'yhat': 'yhat_'+g})

    # Calculate Mean Absolute Error for each state and append to a list
    metric_df = forecast.set_index('ds')[['yhat_'+g]].join(group.set_index('ds').y).reset_index()
    metric_df.dropna(inplace=True)
    metric_df[metric_df['yhat_'+g] < 0] = 0
    #print(metric_df.tail())
    Error = mean_absolute_error(metric_df['y'], metric_df['yhat_'+g])
    MAE.append(Error)

    #Create dataframe with each state's forecasted values
    deaths_project_df = pd.merge(deaths_project_df, forecast.set_index('ds'), how='outer', left_index=True, right_index=True)
```

Calculate 5-day moving average of each state’s positive cases

Build Prophet Model

Add positive cases as a regressor value in prophet model to predict deaths 3 weeks in the future

Calculate MAE for past forecasted values

Create data frame with each state’s forecasted Covid-19 deaths

Figure 9: Components of Prophet Model in Python to predict Covid-19 related deaths

Lastly, the results were smoothed using a 7-day moving average. Due to how deaths are reported, the graphs are often oscillating with the crest occurring mid-week and trough on the weekends. Figure 10 shows the forecasting results before and after smoothing.

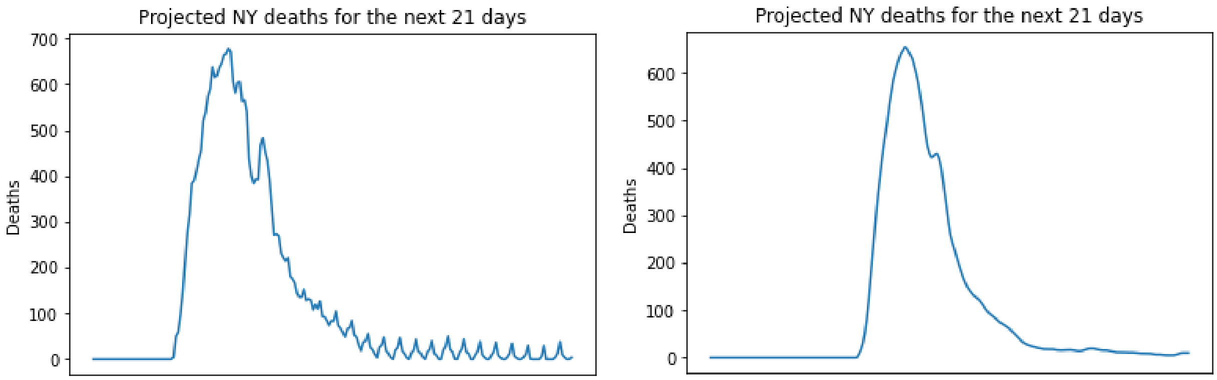


Figure 10. The left and right graph show the Prophet forecasting results for the state of NY before and after smoothing respectively

Covid-19 Deaths - Forecasting Results

These results are best viewed on an interactive dashboard where the end user can customize the view specific states and regions at specific times. The Prophet results in Python are exported to a public Tableau dashboard via a Google Sheets API.

Figure 11 shows an overview of the Covid-19 related deaths which was most recently refreshed on September 13, 2020. At this time, several states had an initial spike in Covid-19 deaths during the Spring of 2020 followed by a sharp decline in the following months while other states were beginning their spike during the Summer of 2020.

Another advantage of Tableau is the ability to pre-select specific values to highlight certain findings in the results. Figure 12 shows how states in the Northeast and Midwest saw an early spike in Covid-19 related deaths around April. At the time of this analysis, deaths declined sharply in the summer and the model predicted that deaths remain low going until October. Conversely, Figure 13 shows that states in the Sun Belt were initially low and started to spike in the Summer months. The model predicted that deaths remain high going into October. The model's Mean Absolute Error (MAE) of past forecasted values is 5.89 deaths per day.

The model is far from perfect and can be further enhanced to better predict Covid-19 related deaths. As society better understands the nature of the virus, additional predictors such as weather, health policy, and demographic features such as population and population density can be added as additional regressors to the model. The prophet model also continues to better predict deaths as it collects more historical data for use in its predictive analysis. In essence, data science is a dynamic discipline and will only continue to advance as more data is gathered and data mining techniques are improved.

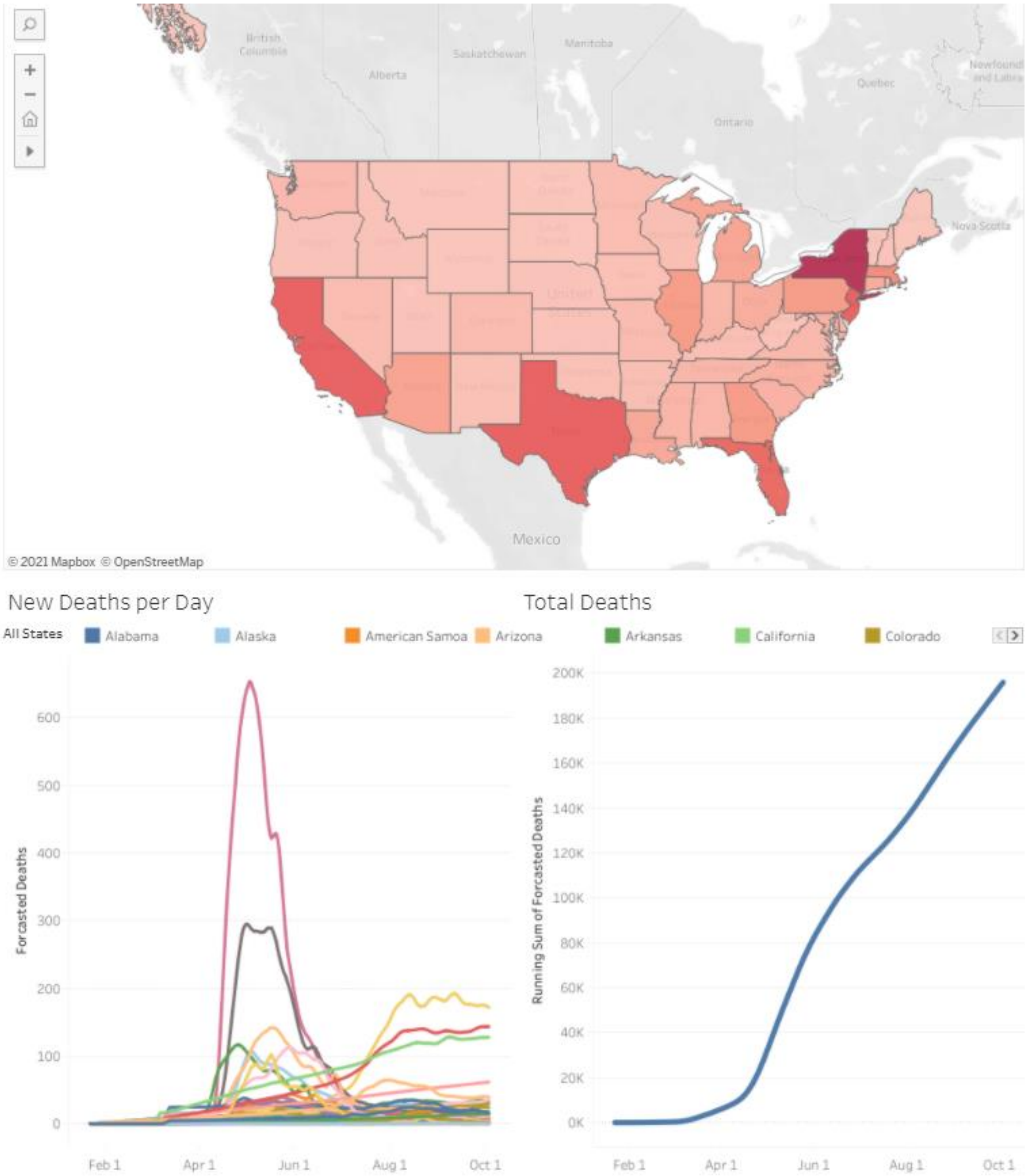


Figure 11. Overview of Prophet Results on Tableau. The map displays the Covid-19 related deaths for each state. The bottom-left line graph displays the time series forecasting for all states while the bottom-right lien graph shows the running tally of total Covid-19 related deaths in the United States.

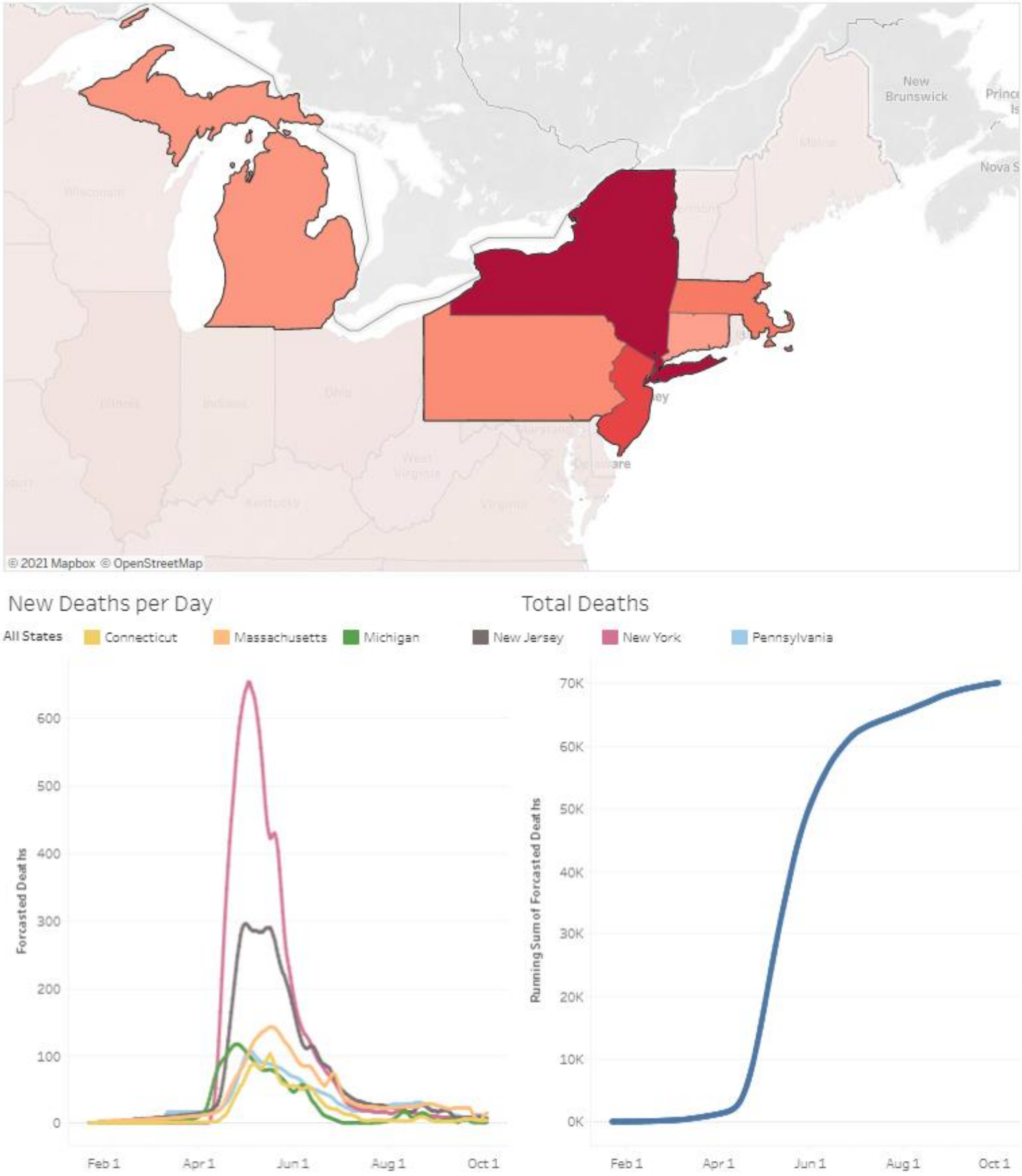


Figure 12. Prophet Results on Tableau of states in the Northeast and Midwest.

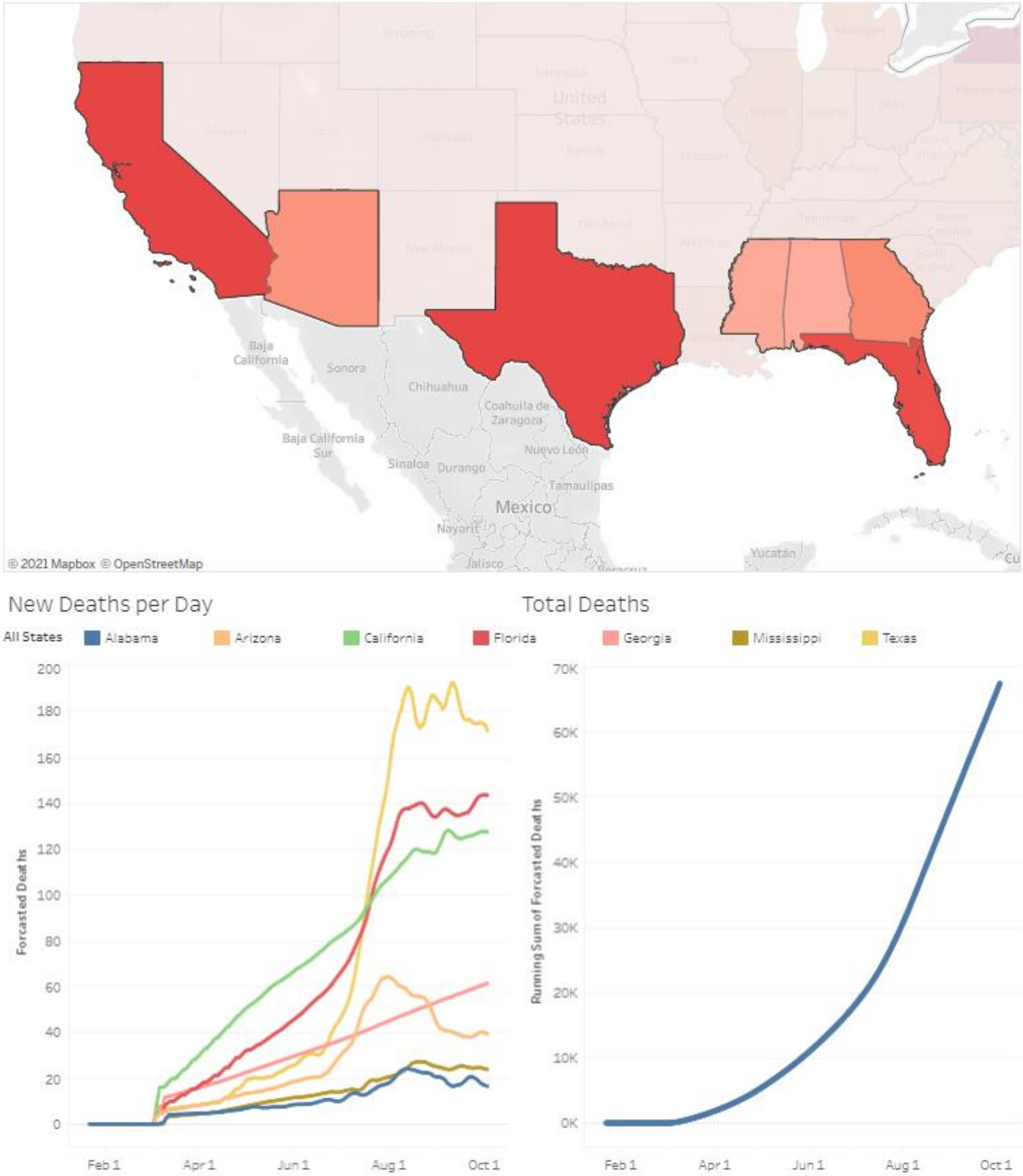


Figure 13. Prophet Results on Tableau of states in the Sun Belt.

Covid-19 Deaths - Linkage to Program Goals

Describe a broad overview of the major practice areas of data science – Data science can provide both descriptive and predictive analytics. Specifically, it can provide an interpretation of historical data or it can use this data to predict future outcomes. These kinds of results are one of the key aspects of data science and both types of analytics are demonstrated in the Covid-19 project. The time series show how past trends look like visually on a time series plot. Additionally, it shows how future trends may look like based on this historical data. Understanding past, present, and future data is particularly useful in leading to impactful decision making.

Collect and organize data – Like the flight data, the Covid-19 death data needed several data cleansing and wrangling steps for the time series analysis to be conducted. Furthermore, this project shows how time series results are best organized as a dashboard for the end-user to review. It allows the user to interact with the data to determine the underlying issues or filter based on their specific needs. Sometimes the end-user just wants to review specific states or certain time periods and an interactive dashboard can allow the user to do exactly that. This also **demonstrates communication skills regarding data and its analysis for managers, IT professionals, programmers, statisticians, and other relevant professionals in their organization.** An interactive dashboard is more suitable for a broader audience as it displays results in a more intuitive form through color-coded maps and line graphs**.** Sometimes the end-user is more focused on the results rather than the analysis. Sometimes the end-user is not savvy when it comes to the technical components of data mining techniques.

This connection from a Python Notebook to a Tableau dashboard also allows the data to be refreshed continuously. Data is expanding in real-time, and an active connection is generally required for models to continuously learn and produce better outputs.

Identify patterns in data via visualization, statistical analysis, and data mining - The results in the time series forecasting provide what is perhaps the most advantageous feature of predictive analytics: using historical patterns to predict future patterns and outcomes. This project demonstrated how the model utilized past trends in Covid-19 deaths and infection rates to predict the trajectory of the pandemic in the coming months.

Develop a plan of action to implement the business decisions derived from the analyses - Covid-19 has created a global health crisis unlike any other in recent memory and understanding the virus can help society better respond, acclimate, and curb the negative effects of this ongoing crisis. The results can describe how society is responding to the pandemic and what the foreseeable future will look like. Is society turning the corner, or will Covid-19 related deaths continue to be at high levels? The Covid-19 predictions show how data can be used to identify patterns and develop an action plan. If the predictive analytics show that the numbers aren't changing, what more does society need to do in terms of vaccinations, health policy, and awareness to curb the negative effects of the ongoing crisis?

IST 652 – Scripting for Data Analysis

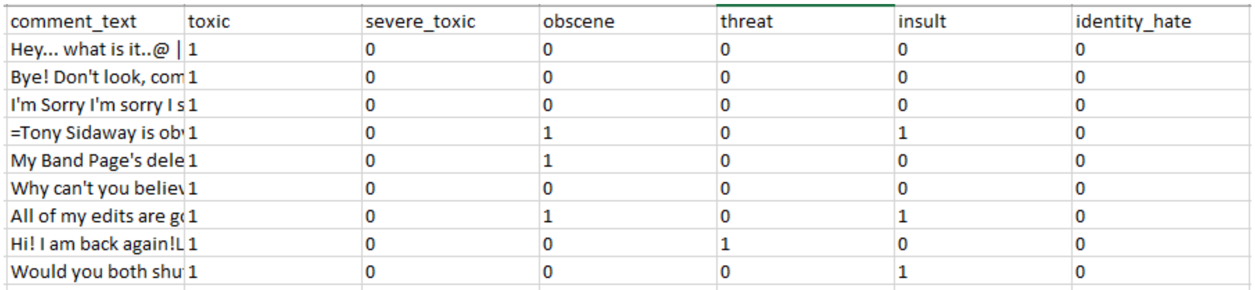
One of the more recent trends in data science is the use of deep learning to create models for natural language processing (NLP). These models have been particularly popular as they can often produce superior results when compared to other learning algorithms. The use of deep learning is demonstrated in a Toxic Comment Classification project for IST 652.

Toxic Comments – Introduction

Offensive and vulgar language have plagued internet communities since its inception. When considering the massive scale in which humans are interacting and communicating with each other on the web, it is nearly impossible to censor toxic comments through manual human labor. To demonstrate how machine learning algorithms can help, a neural network model is trained to identify and classify text data based on its toxicity.

Toxic Comments – Neural Networks

The data comes from a Kaggle Competition where contestants are challenged to create a model that predicts a text's category as either toxic, severely toxic, obscene, threatening, insulting, or identity hate.



comment_text	toxic	severe_toxic	obscene	threat	insult	identity_hate
Hey... what is it..@ |	1	0	0	0	0	0
Bye! Don't look, com	1	0	0	0	0	0
I'm Sorry I'm sorry I s	1	0	0	0	0	0
=Tony Sidaway is ob	1	0	1	0	1	0
My Band Page's dele	1	0	1	0	0	0
Why can't you believ	1	0	0	0	0	0
All of my edits are g	1	0	1	0	1	0
Hi! I am back again!L	1	0	0	1	0	0
Would you both shu	1	0	0	0	1	0

Figure 14. Sample of data from the Toxic Comment Classification Challenge. Each comment has been scored by a human based on its toxicity.

TenserFlow 2.0 was utilized to create a long short-term memory model (LSTM). This is a type of recurrent neural network and are very powerful in recognizing patterns and solving sequence prediction problems especially for NLP.

Just like other data analysis methods, NLP requires extensive data collection, organization, and manipulation to allow for proper analysis. In this example, the text data is first tokenized so that they may be used as features in the model. They must also be padded to a fixed length. Figure 15 shows the distribution of word lengths which is used to estimate the optimal amount of padding. In this case, 300 was determined to be the maximum length.

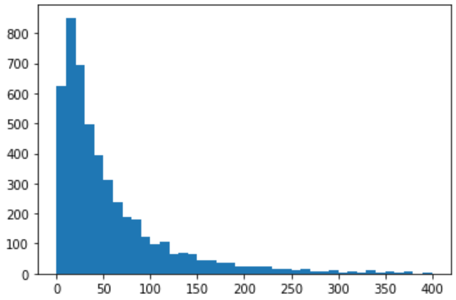


Figure 15. Distribution of word lengths in the Toxic Comments dataset

Unlike the unsupervised learning techniques used previously in the flight dataset, neural networks are a supervised technique which involve training a model with labelled data. To do this, 3,500 comments are used as test data and 1,500 comments are used as training data.

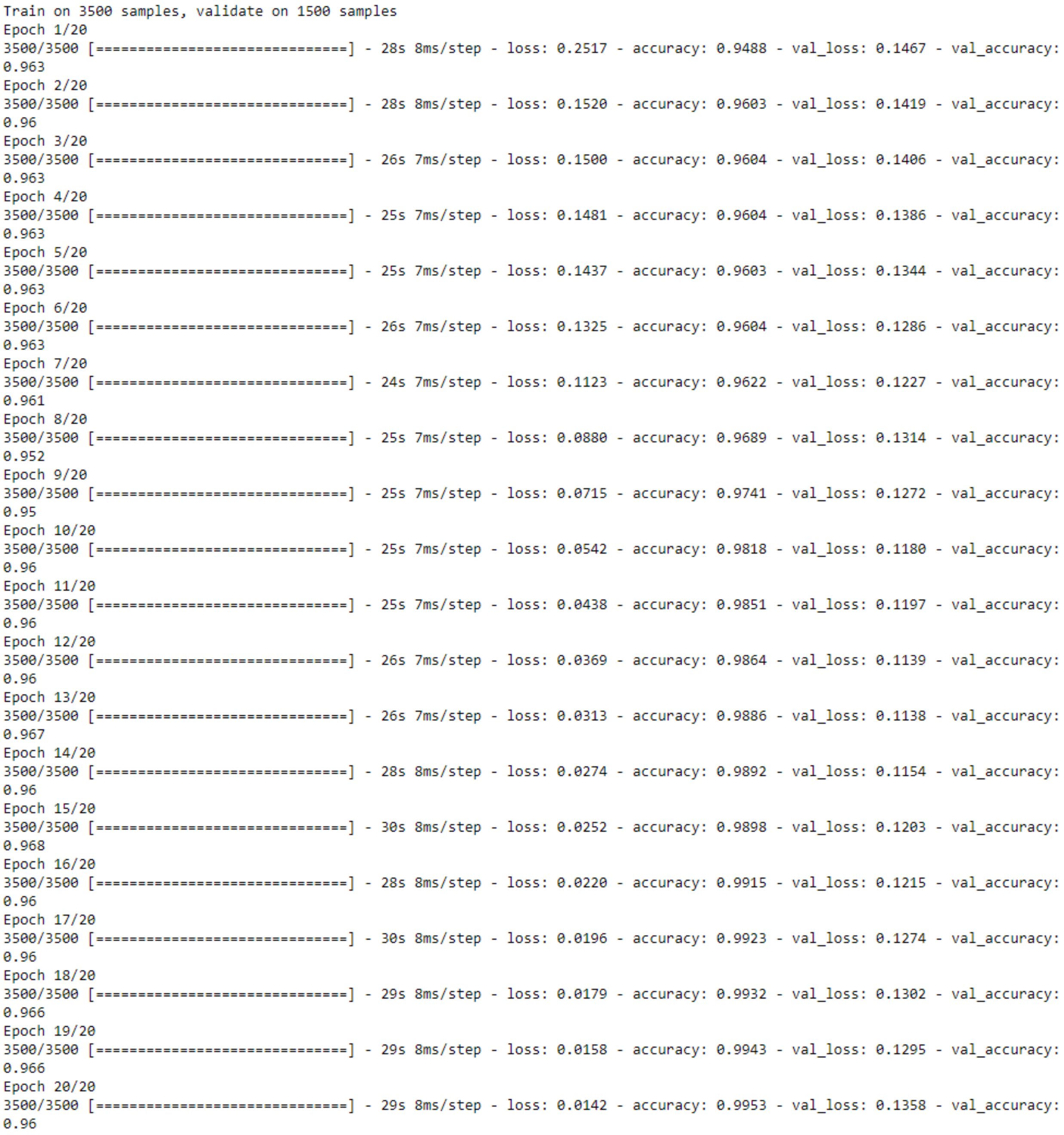
One of the hallmark characteristics of building a LSTM network is the ability to select the parameters or layers in the model. In this example, the following layers were selected:

- **Embedding**: Important for natural language processing as it helps reduces the dimensionality of discrete variables
- **LSTM**: Adds sequence-prediction and neural-network capabilities
- **GlobalMaxPool**: Reduces over-fitting
- **Dense**: A dense layer is a fully connected layer where each 'neuron' is connected to all other variables from the next layer.

```
inp = Input(shape=(maxlen, ))
embed_size = 128
x = Embedding(max_features, embed_size)(inp)
x = LSTM(60, return_sequences=True,name='lstm_layer')(x)
x = GlobalMaxPool1D()(x)
x = Dense(6, activation="sigmoid")(x)
model = Model(inputs=inp, outputs=x)
model.compile(loss='binary_crossentropy',
              optimizer='adam',
              metrics=['accuracy'])
```

Figure 16. Code snippet of LSTM model using the defined layers

The data is then input to the model to produce the results shown in Figure 17.



```
Train on 3500 samples, validate on 1500 samples
Epoch 1/20
3500/3500 [==============================] - 28s 8ms/step - loss: 0.2517 - accuracy: 0.9488 - val_loss: 0.1467 - val_accuracy:
0.963
Epoch 2/20
3500/3500 [==============================] - 28s 8ms/step - loss: 0.1520 - accuracy: 0.9603 - val_loss: 0.1419 - val_accuracy:
0.96
Epoch 3/20
3500/3500 [==============================] - 26s 7ms/step - loss: 0.1500 - accuracy: 0.9604 - val_loss: 0.1406 - val_accuracy:
0.963
Epoch 4/20
3500/3500 [==============================] - 25s 7ms/step - loss: 0.1481 - accuracy: 0.9604 - val_loss: 0.1386 - val_accuracy:
0.963
Epoch 5/20
3500/3500 [==============================] - 25s 7ms/step - loss: 0.1437 - accuracy: 0.9603 - val_loss: 0.1344 - val_accuracy:
0.963
Epoch 6/20
3500/3500 [==============================] - 26s 7ms/step - loss: 0.1325 - accuracy: 0.9604 - val_loss: 0.1286 - val_accuracy:
0.963
Epoch 7/20
3500/3500 [==============================] - 24s 7ms/step - loss: 0.1123 - accuracy: 0.9622 - val_loss: 0.1227 - val_accuracy:
0.961
Epoch 8/20
3500/3500 [==============================] - 25s 7ms/step - loss: 0.0880 - accuracy: 0.9689 - val_loss: 0.1314 - val_accuracy:
0.952
Epoch 9/20
3500/3500 [==============================] - 25s 7ms/step - loss: 0.0715 - accuracy: 0.9741 - val_loss: 0.1272 - val_accuracy:
0.95
Epoch 10/20
3500/3500 [==============================] - 25s 7ms/step - loss: 0.0542 - accuracy: 0.9818 - val_loss: 0.1180 - val_accuracy:
0.96
Epoch 11/20
3500/3500 [==============================] - 25s 7ms/step - loss: 0.0438 - accuracy: 0.9851 - val_loss: 0.1197 - val_accuracy:
0.96
Epoch 12/20
3500/3500 [==============================] - 26s 7ms/step - loss: 0.0369 - accuracy: 0.9864 - val_loss: 0.1139 - val_accuracy:
0.96
Epoch 13/20
3500/3500 [==============================] - 26s 7ms/step - loss: 0.0313 - accuracy: 0.9886 - val_loss: 0.1138 - val_accuracy:
0.967
Epoch 14/20
3500/3500 [==============================] - 28s 8ms/step - loss: 0.0274 - accuracy: 0.9892 - val_loss: 0.1154 - val_accuracy:
0.96
Epoch 15/20
3500/3500 [==============================] - 30s 8ms/step - loss: 0.0252 - accuracy: 0.9898 - val_loss: 0.1203 - val_accuracy:
0.968
Epoch 16/20
3500/3500 [==============================] - 28s 8ms/step - loss: 0.0220 - accuracy: 0.9915 - val_loss: 0.1215 - val_accuracy:
0.96
Epoch 17/20
3500/3500 [==============================] - 30s 8ms/step - loss: 0.0196 - accuracy: 0.9923 - val_loss: 0.1274 - val_accuracy:
0.96
Epoch 18/20
3500/3500 [==============================] - 29s 8ms/step - loss: 0.0179 - accuracy: 0.9932 - val_loss: 0.1302 - val_accuracy:
0.966
Epoch 19/20
3500/3500 [==============================] - 29s 8ms/step - loss: 0.0158 - accuracy: 0.9943 - val_loss: 0.1295 - val_accuracy:
0.966
Epoch 20/20
3500/3500 [==============================] - 29s 8ms/step - loss: 0.0142 - accuracy: 0.9953 - val_loss: 0.1358 - val_accuracy:
0.96
```

Figure 17. LSTM Model Results for Toxic Comments Classification

Each epoch indicates "one pass over the entire dataset" and is used as a marker to separate the training into distinct phases. As the model changes over time, the 'loss' and 'val_loss' decreases which indicates the "average of the losses" over each batch of data is decreasing. 'loss' is applied to the train set while 'val_loss' is for the test or validation set. Conversely, the 'acc' and 'val_acc' is increasing over time which represents the percentage of instances that are correctly classified. 'acc' applies to the train set while the 'val_acc' is for the test or validation set. Overall, the model appears to have trained successfully and continued to improve over time.

Toxic Comments - Linkage to Program Goals

Identify patterns in data via visualization, statistical analysis, and data mining - The results show how data science can be used to Identify patterns in data through data mining. In this case, the LSTM model was able to predict how text can lead to its toxicity classification based on patterns found in the training data. This model is most useful to managers and IT professionals who wish to identify ways to automatically moderate submitted text comments. It can be seen from the results that neural networks are able to predict comment toxicity with very high accuracy.

IST 659 Data Admin Concepts & Database Management

The topic of privacy and protection of identifiable health information is one of the most discussed controversial topics in data science. While there is certainly a need to protect patients and their health information, some protected health information (PHI) is helpful and sometimes even necessary for use in medical research and clinical trials. The Health Insurance Portability and Accountability Act (HIPAA) privacy rule sets forth policies to protect such information and defines 18 identifiers such as name, address, SSN numbers, and even finger or voice prints as protected health information (PHI) which must be “de-identified” from a dataset in the use of any communication.

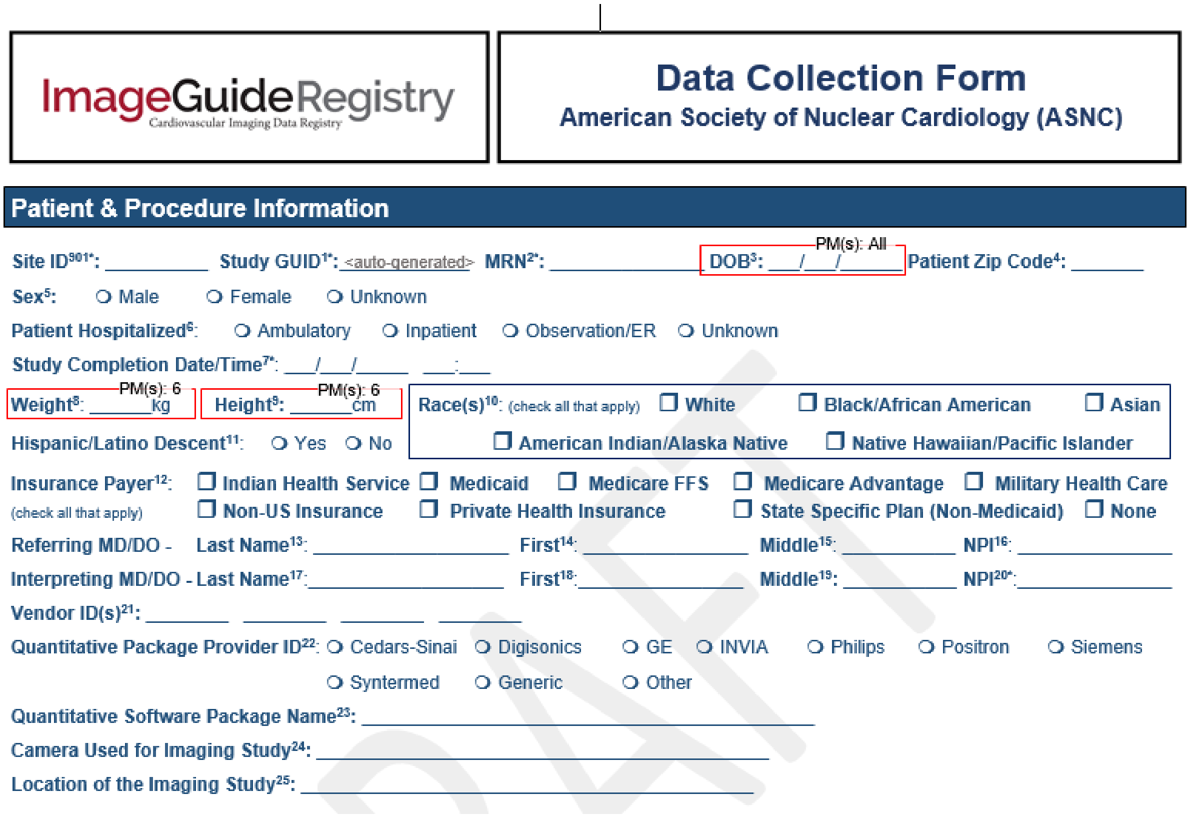
The collection de-identified data is demonstrated in an IST 659 final project where a database is designed and created for the collection of cardiology medical reports.

Nuclear Cardiology Database

In this example, a database is designed for a clinical data registry which collects and stores medical data. This database focuses on the collection of nuclear stress report data which can then be used to answer specific questions with the goal of improving the quality of health care in nuclear cardiology.

A nuclear stress test is a noninvasive imaging technique in assessing blood flow to diagnose heart disease and guide treatment of disorders. During the test, a radioactive tracer is injected into the patient which can then be detected by a special camera to produce images of the heart. After a hospital or outpatient center completes a nuclear stress test, the patient and procedure information is recorded in a medical report which is then sent to the clinical data registry for analysis.

A typical medical report is filled with PHI. As shown in Figure 18 of a sample nuclear cardiology medical report, many fields such as the medical record number (MRN), date of birth (DOB), and patient zip code are considered PHI and must be de-identified.



ImageGuideRegistry
Cardiovascular Imaging Data Registry

Data Collection Form
American Society of Nuclear Cardiology (ASNC)

Patient & Procedure Information

Site ID[901*]: __________ Study GUID[1*]: <auto-generated> MRN[2*]: ______________ DOB[3]: PM(s): All ___/___/______ Patient Zip Code[4]: _______

Sex[5]: ○ Male ○ Female ○ Unknown

Patient Hospitalized[6]: ○ Ambulatory ○ Inpatient ○ Observation/ER ○ Unknown

Study Completion Date/Time[7*]: ___/___/_____ ___:___

Weight[8]: PM(s): 6 _______kg Height[9]: PM(s): 6 _______cm Race(s)[10]: (check all that apply) □ White □ Black/African American □ Asian □ American Indian/Alaska Native □ Native Hawaiian/Pacific Islander

Hispanic/Latino Descent[11]: ○ Yes ○ No

Insurance Payer[12]: (check all that apply) □ Indian Health Service □ Medicaid □ Medicare FFS □ Medicare Advantage □ Military Health Care □ Non-US Insurance □ Private Health Insurance □ State Specific Plan (Non-Medicaid) □ None

Referring MD/DO - Last Name[13]: __________________ First[14]: ________________ Middle[15]: ___________ NPI[16]: ______________

Interpreting MD/DO - Last Name[17]: __________________ First[18]: ________________ Middle[19]: ___________ NPI[20*]: ______________

Vendor ID(s)[21]: ________ ________ ________ ________

Quantitative Package Provider ID[22]: ○ Cedars-Sinai ○ Digisonics ○ GE ○ INVIA ○ Philips ○ Positron ○ Siemens ○ Syntermed ○ Generic ○ Other

Quantitative Software Package Name[23]: ____

Camera Used for Imaging Study[24]: ____

Location of the Imaging Study[25]: ____

Figure 18. Sample of medical report form that is collected in the ImageGuide Registry

To focus on areas that can be used for research and quality improvement, the database extracts the fields defined in the data dictionary in Figure 19.

Entity	Attribute	Properties
Study	StudyID	Required and Unique
	StudyDate	Required. Indicates when study is performed.
	StressHR	Indicates a patient's heart rate when stressed during the study

Indications	Indications	Required. Describes the patient’s conditions that led to the nuclear stress test
	StudyAppropriateness	Derived. Describes if a nuclear study is appropriate or inappropriate based on indications
InterpretingPhysician	Name	Required – First and Last Name
	NPI (National Provider Identifier)	Required and Unique
Practice	PracticeName	Required
	PracticeAddress	Required
ImagingProtocol	ImagingProtocol	Required. Describes if imaging is done when patient is on stress, rest, or both stress and rest.
StressType	StressType	Describes if a patient was stressed through exercise or through a pharmacologic agent
StressRadiopharmaceutical	StressRadiopharmaceutical	Describes type of tracer used during the stress portion of the test
RestRadiopharmaceutical	RestRadiopharmaceutical	Describes type of tracer used during the rest portion of the test

Figure 19. Data Dictionary of entities and attributes

The collected data will be normalized to minimize redundancy and optimize data structures. To create a logical model, the following considerations were made:

Relations: For each nuclear stress test, there is one interpreting physician, practice location, radio pharmaceuticals, stress type, and protocol. Therefore, the model indicates a “Many-To-One" relationship between the study entity and these entities. For indications, because this is a “Many-to Many” relationship with the study entity, an intermediate Study-Indications List entity is created.

Normalization: The relations in this model are in third normal form as they do not contain any functional or transitive dependencies. There are no functional dependencies that need to be corrected since each study has a single attribute key (StudyID). All transitive dependencies were removed by creating a new entity with non-prime attributes that are dependent on other non-key attributes.

Data Types: The varchar data type is used for most attributes with the exception of StudyDate, the Physician's National Provider Identifier (NPI), and stress Heart Rate. The data type has been indicated for each attribute in the logical model.

Required Fields: The physician information, practice information, indications, and protocol are required fields for a study.

This database's entities and their relationships are graphically illustrated by the Normalized Logical Model shown in Figure 20.

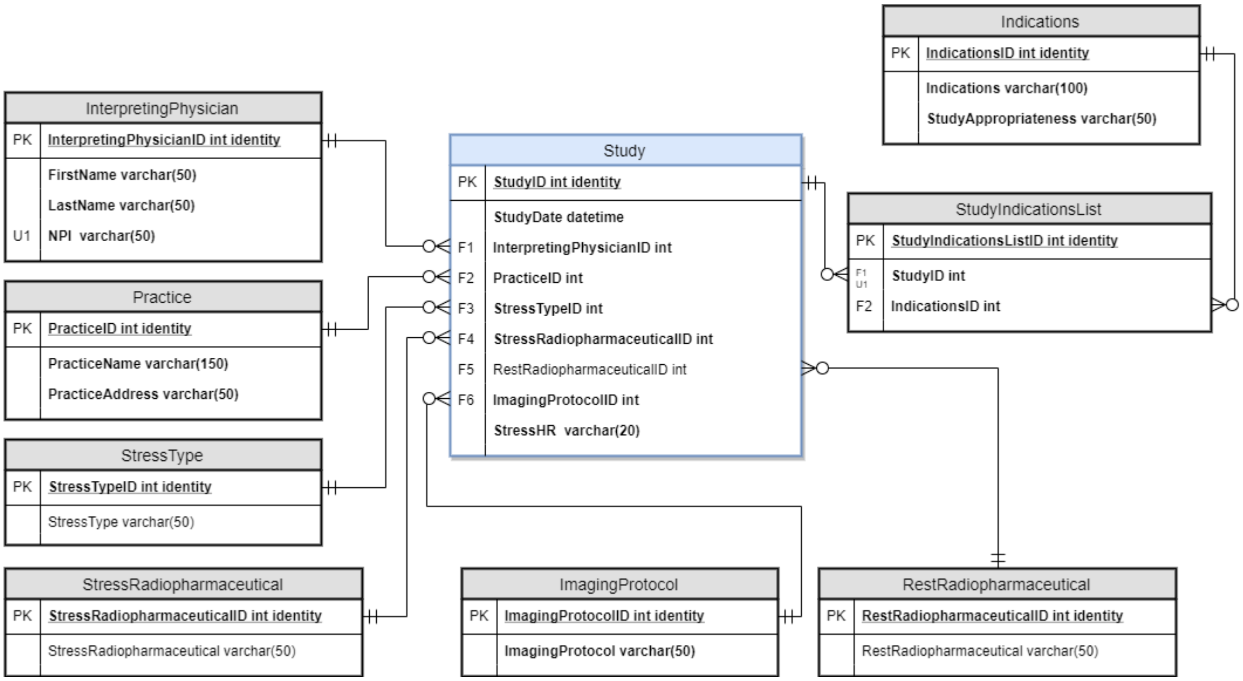
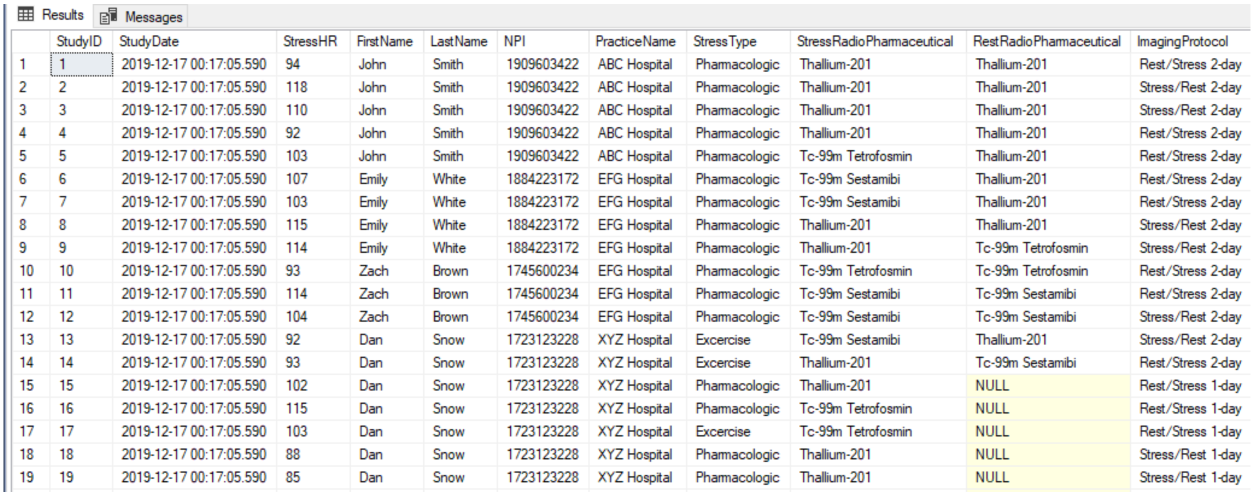

Indications
PK IndicationsID int identity
Indications varchar(100)
StudyAppropriateness varchar(50)
InterpretingPhysician
PK InterpretingPhysicianID int identity
FirstName varchar(50)
LastName varchar(50)
U1 NPI varchar(50)
Study
PK StudyID int identity
StudyDate datetime
F1 InterpretingPhysicianID int
F2 PracticeID int
F3 StressTypeID int
F4 StressRadiopharmaceuticalID int
F5 RestRadiopharmaceuticalID int
F6 ImagingProtocolID int
StressHR varchar(20)
StudyIndicationsList
PK StudyIndicationsListID int identity
F1 U1 StudyID int
F2 IndicationsID int
Practice
PK PracticeID int identity
PracticeName varchar(150)
PracticeAddress varchar(50)
StressType
PK StressTypeID int identity
StressType varchar(50)
StressRadiopharmaceutical
PK StressRadiopharmaceuticalID int identity
StressRadiopharmaceutical varchar(50)
ImagingProtocol
PK ImagingProtocolID int identity
ImagingProtocol varchar(50)
RestRadiopharmaceutical
PK RestRadiopharmaceuticalID int identity
RestRadiopharmaceutical varchar(50)

Figure 20. Database entity relationship diagram for a nuclear cardiology data registry

Now that the database has been designed, the tables can then be coded in SQL. Sample data can also be inserted into the database which is shown in Figure 21.



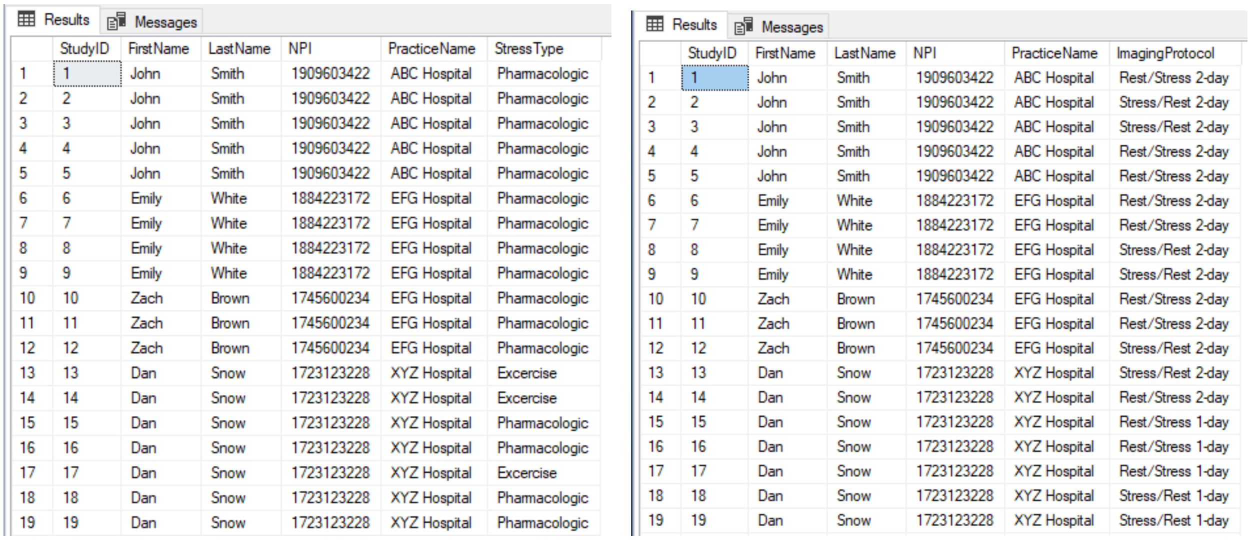
	StudyID	StudyDate	StressHR	FirstName	LastName	NPI	PracticeName	StressType	StressRadioPhamaceutical	RestRadioPhamaceutical	ImagingProtocol
1	1	2019-12-17 00:17:05.590	94	John	Smith	1909603422	ABC Hospital	Phamacologic	Thallium-201	Thallium-201	Rest/Stress 2-day
2	2	2019-12-17 00:17:05.590	118	John	Smith	1909603422	ABC Hospital	Phamacologic	Thallium-201	Thallium-201	Stress/Rest 2-day
3	3	2019-12-17 00:17:05.590	110	John	Smith	1909603422	ABC Hospital	Phamacologic	Thallium-201	Thallium-201	Stress/Rest 2-day
4	4	2019-12-17 00:17:05.590	92	John	Smith	1909603422	ABC Hospital	Phamacologic	Thallium-201	Thallium-201	Rest/Stress 2-day
5	5	2019-12-17 00:17:05.590	103	John	Smith	1909603422	ABC Hospital	Phamacologic	Tc-99m Tetrofosmin	Thallium-201	Rest/Stress 2-day
6	6	2019-12-17 00:17:05.590	107	Emily	White	1884223172	EFG Hospital	Phamacologic	Tc-99m Sestamibi	Thallium-201	Rest/Stress 2-day
7	7	2019-12-17 00:17:05.590	103	Emily	White	1884223172	EFG Hospital	Phamacologic	Tc-99m Sestamibi	Thallium-201	Rest/Stress 2-day
8	8	2019-12-17 00:17:05.590	115	Emily	White	1884223172	EFG Hospital	Phamacologic	Thallium-201	Thallium-201	Stress/Rest 2-day
9	9	2019-12-17 00:17:05.590	114	Emily	White	1884223172	EFG Hospital	Phamacologic	Thallium-201	Tc-99m Tetrofosmin	Stress/Rest 2-day
10	10	2019-12-17 00:17:05.590	93	Zach	Brown	1745600234	EFG Hospital	Phamacologic	Tc-99m Tetrofosmin	Tc-99m Tetrofosmin	Rest/Stress 2-day
11	11	2019-12-17 00:17:05.590	114	Zach	Brown	1745600234	EFG Hospital	Phamacologic	Tc-99m Sestamibi	Tc-99m Sestamibi	Rest/Stress 2-day
12	12	2019-12-17 00:17:05.590	104	Zach	Brown	1745600234	EFG Hospital	Phamacologic	Tc-99m Sestamibi	Tc-99m Sestamibi	Stress/Rest 2-day
13	13	2019-12-17 00:17:05.590	92	Dan	Snow	1723123228	XYZ Hospital	Excercise	Tc-99m Sestamibi	Thallium-201	Stress/Rest 2-day
14	14	2019-12-17 00:17:05.590	93	Dan	Snow	1723123228	XYZ Hospital	Excercise	Thallium-201	Tc-99m Sestamibi	Rest/Stress 2-day
15	15	2019-12-17 00:17:05.590	102	Dan	Snow	1723123228	XYZ Hospital	Phamacologic	Thallium-201	NULL	Rest/Stress 1-day
16	16	2019-12-17 00:17:05.590	115	Dan	Snow	1723123228	XYZ Hospital	Phamacologic	Tc-99m Tetrofosmin	NULL	Rest/Stress 1-day
17	17	2019-12-17 00:17:05.590	103	Dan	Snow	1723123228	XYZ Hospital	Excercise	Tc-99m Tetrofosmin	NULL	Rest/Stress 1-day
18	18	2019-12-17 00:17:05.590	88	Dan	Snow	1723123228	XYZ Hospital	Phamacologic	Thallium-201	NULL	Stress/Rest 1-day
19	19	2019-12-17 00:17:05.590	85	Dan	Snow	1723123228	XYZ Hospital	Phamacologic	Thallium-201	NULL	Stress/Rest 1-day

Figure 21. Sample of SQL database with de-identified medical record data

These elements in this database can answers several questions; all of which are important in finding opportunities for quality improvement in the cardiology specialty.

- What is the percentage of patients who are stressed through exercise vs. through a pharmacologic agent? Stress through exercise generally delivers more prognostic information.
- What is the percentage of nuclear studies that are still using Thallium? Thallium is very high in radiation, outdated, and not recommended for use in nuclear stress tests.
- How many patients are undergoing a 2-day protocol vs. a 1-day protocol? 1-day protocols are generally recommended as they involve lower radiations.
- Which physicians and hospitals/outpatient centers are following recommended guidelines?

The database can be quickly and seamlessly queried to answer these specific data questions. Figure 22 displays examples of query results.



	StudyID	FirstName	LastName	NPI	PracticeName	StressType
1	1	John	Smith	1909603422	ABC Hospital	Pharmacologic
2	2	John	Smith	1909603422	ABC Hospital	Pharmacologic
3	3	John	Smith	1909603422	ABC Hospital	Pharmacologic
4	4	John	Smith	1909603422	ABC Hospital	Pharmacologic
5	5	John	Smith	1909603422	ABC Hospital	Pharmacologic
6	6	Emily	White	1884223172	EFG Hospital	Pharmacologic
7	7	Emily	White	1884223172	EFG Hospital	Pharmacologic
8	8	Emily	White	1884223172	EFG Hospital	Pharmacologic
9	9	Emily	White	1884223172	EFG Hospital	Pharmacologic
10	10	Zach	Brown	1745600234	EFG Hospital	Pharmacologic
11	11	Zach	Brown	1745600234	EFG Hospital	Pharmacologic
12	12	Zach	Brown	1745600234	EFG Hospital	Pharmacologic
13	13	Dan	Snow	1723123228	XYZ Hospital	Excercise
14	14	Dan	Snow	1723123228	XYZ Hospital	Excercise
15	15	Dan	Snow	1723123228	XYZ Hospital	Pharmacologic
16	16	Dan	Snow	1723123228	XYZ Hospital	Pharmacologic
17	17	Dan	Snow	1723123228	XYZ Hospital	Excercise
18	18	Dan	Snow	1723123228	XYZ Hospital	Pharmacologic
19	19	Dan	Snow	1723123228	XYZ Hospital	Pharmacologic

	StudyID	FirstName	LastName	NPI	PracticeName	ImagingProtocol
1	1	John	Smith	1909603422	ABC Hospital	Rest/Stress 2-day
2	2	John	Smith	1909603422	ABC Hospital	Stress/Rest 2-day
3	3	John	Smith	1909603422	ABC Hospital	Stress/Rest 2-day
4	4	John	Smith	1909603422	ABC Hospital	Rest/Stress 2-day
5	5	John	Smith	1909603422	ABC Hospital	Rest/Stress 2-day
6	6	Emily	White	1884223172	EFG Hospital	Rest/Stress 2-day
7	7	Emily	White	1884223172	EFG Hospital	Rest/Stress 2-day
8	8	Emily	White	1884223172	EFG Hospital	Stress/Rest 2-day
9	9	Emily	White	1884223172	EFG Hospital	Stress/Rest 2-day
10	10	Zach	Brown	1745600234	EFG Hospital	Rest/Stress 2-day
11	11	Zach	Brown	1745600234	EFG Hospital	Rest/Stress 2-day
12	12	Zach	Brown	1745600234	EFG Hospital	Stress/Rest 2-day
13	13	Dan	Snow	1723123228	XYZ Hospital	Stress/Rest 2-day
14	14	Dan	Snow	1723123228	XYZ Hospital	Rest/Stress 2-day
15	15	Dan	Snow	1723123228	XYZ Hospital	Rest/Stress 1-day
16	16	Dan	Snow	1723123228	XYZ Hospital	Rest/Stress 1-day
17	17	Dan	Snow	1723123228	XYZ Hospital	Rest/Stress 1-day
18	18	Dan	Snow	1723123228	XYZ Hospital	Stress/Rest 1-day
19	19	Dan	Snow	1723123228	XYZ Hospital	Stress/Rest 1-day

Figure 22. Sample of SQL query results that answer specific data questions

Nuclear Cardiology Database – Linkage to Program Goals

Demonstrate communication skills regarding data and its analysis for managers, IT professionals, programmers, statisticians, and other relevant professionals in their organization. - The database queries are useful to several stakeholders including health practices who want to determine what areas of nuclear cardiology need improvement. This is also of interest to health insurance and other payors who are interested in knowing the rate of appropriate and inappropriate testing. Finally, and most importantly, it is important for patient health and treatment. Nuclear cardiology includes levels of radiation, risk, and adverse effects, and maintaining a high quality in this practice is important in the cardiovascular field to make safe and proper diagnoses.

Synthesize the ethical dimensions of data science practice – Data science is not without its ethical limits. While personal data can be useful in gaining actionable insights, scientists must be careful not to invade privacy. In the case of medical research, privacy rules for protected health information must be considered to protect patient privacy. This project demonstrates how a database for medical research can be created that strictly de-identifies all PHI fields. The database is in full compliance with HIPAA privacy rules for the protection of all individually identifiable health information.

Conclusion

The projects described in this report demonstrate how the variety of courses offered in the Applied Data Science Program at Syracuse University link to several key concepts in the field. Several areas in data science including data collection, management, analysis, and communication for decision-making were discussed at length and synthesized to demonstrate the overall learning throughout the program.

The learning goals include the collection, organizing, and cleansing of data which is demonstrated in the flight project for IST 707 and text data for the IST 652 toxic comments project. It also includes the identification of patterns in data visualizations, analysis, and data mining such as k-means clustering results and time neural networks for NLP. Ethical dimensions of data science were also discussed including the issue of privacy and patient health information when collecting and storing data for research in the medical field. Finally, the projects demonstrate how the analysis and results can be used to communicate skills, findings, and recommendations for action plans and strategies to make improvements to different aspects of society.

Data science is a dynamic field and data mining and analysis techniques will only continue advancing and reaching new heights. While several learning program goals have been demonstrated in this portfolio, the projects are only merely scratching the surface as data science is and will continue to be a lifelong learning mission.