Portfolio Milestone – Draft Submission

Jose Reyes

Contents

[Introduction 2](#_Toc64263835)

[IST 707 - Data Analytics 2](#_Toc64263836)

[Flight Data - K-means clustering results 4](#_Toc64263837)

[Flight Data - Association rules mining results 6](#_Toc64263838)

[Flight Data - Overall Results Analysis 7](#_Toc64263839)

[IST 718 - Big Data Analytics 7](#_Toc64263840)

[Covid-19 Deaths - Time series forecasting using Prophet 7](#_Toc64263841)

[Covid-19 Deaths - Forecasting Results 9](#_Toc64263842)

[Covid-19 Deaths - Overall Results Analysis 13](#_Toc64263843)

[IST 652 – Scripting for Data Analysis 13](#_Toc64263844)

[Toxic Comments – Neural Networks 14](#_Toc64263845)

[Toxic Comments - Overall Results Analysis 16](#_Toc64263846)

[IST 659 Data Admin Concepts & Database Management 16](#_Toc64263847)

[Nuclear Cardiology Database 16](#_Toc64263848)

[Nuclear Cardiology Database – Overall Analysis 20](#_Toc64263849)

[Conclusion 21](#_Toc64263850)

**Project Codes and Supporting Documents can be found on:**

[**https://github.com/JoeR1221/Data-Science-Portfolio**](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 many of these courses focus on different topics from business analytics to machine learning, each one can be described and linked to one of the program’s overall learning goals.

In this portfolio, projects from several different courses will be described, demonstrated, and linked to 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).

# 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, decision trees, and Naïve Bayes classifier. Data mining is the extraction of interesting and previously unknown knowledge through a non-trivial task which is perhaps the hallmark characteristic that separates data mining from more trivial tasks such as a simple query.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 example, a dataset on flight data from the United States Bureau of Transportation Statistics.

Before any analysis can be conducted for any data mining projects, data must be collected and organized into a form that is used for data analysis. This step can be an excruciatingly long and agonizing process, but it can be as important as any other step in the data analysis process.

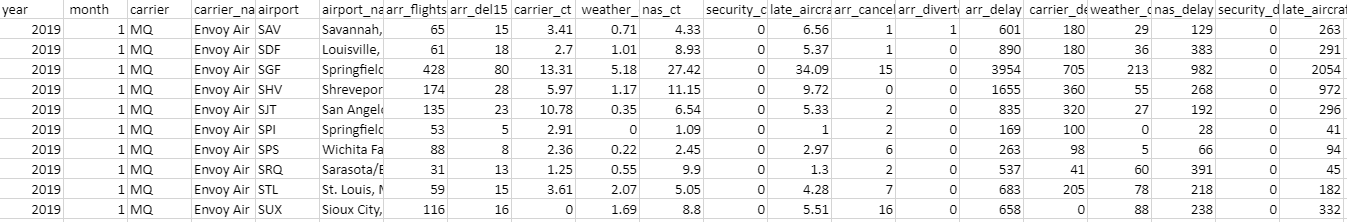


Figure 1. Sample of raw flight data from the United States Bureau of Transportation Statistics

While the raw flight data shown in Figure 1 is already organized into neat rows and columns, it is unusable in its current form for several data mining techniques. The motivation in this project is to identify patterns in flight delays and cancellation among the different airports, airline carriers, and time of year. In order 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.

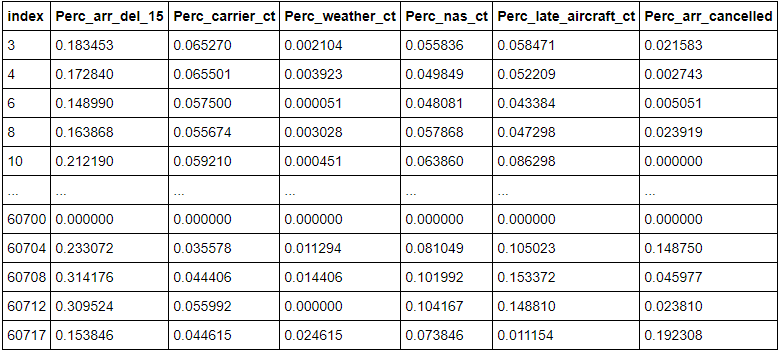


Figure 2. Sample of flight data after data cleansing for k-means clustering

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.

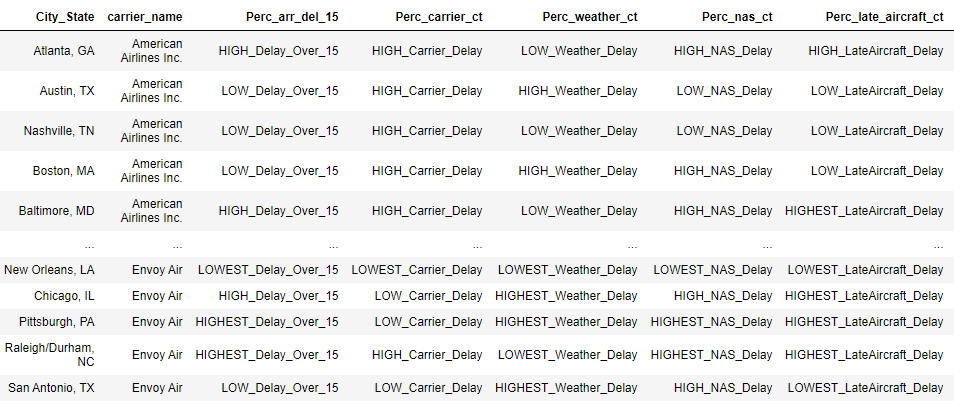


Figure 3. Sample of flight data after data cleansing for association rules mining

## Flight Data - K-means clustering results

There are several ways to visualize the k-means clustering results. The main goal from this analysis 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. The elbow method determined that the data is best organized into 3 different clusters for 6 different features.

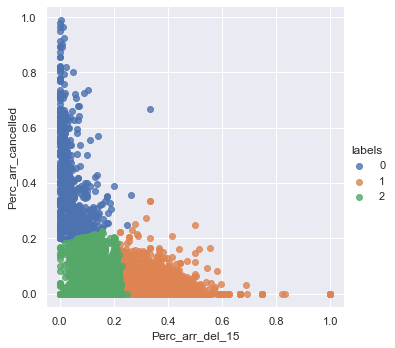


Figure 4. Two-dimensional scatterplot of K-means clustering results

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.

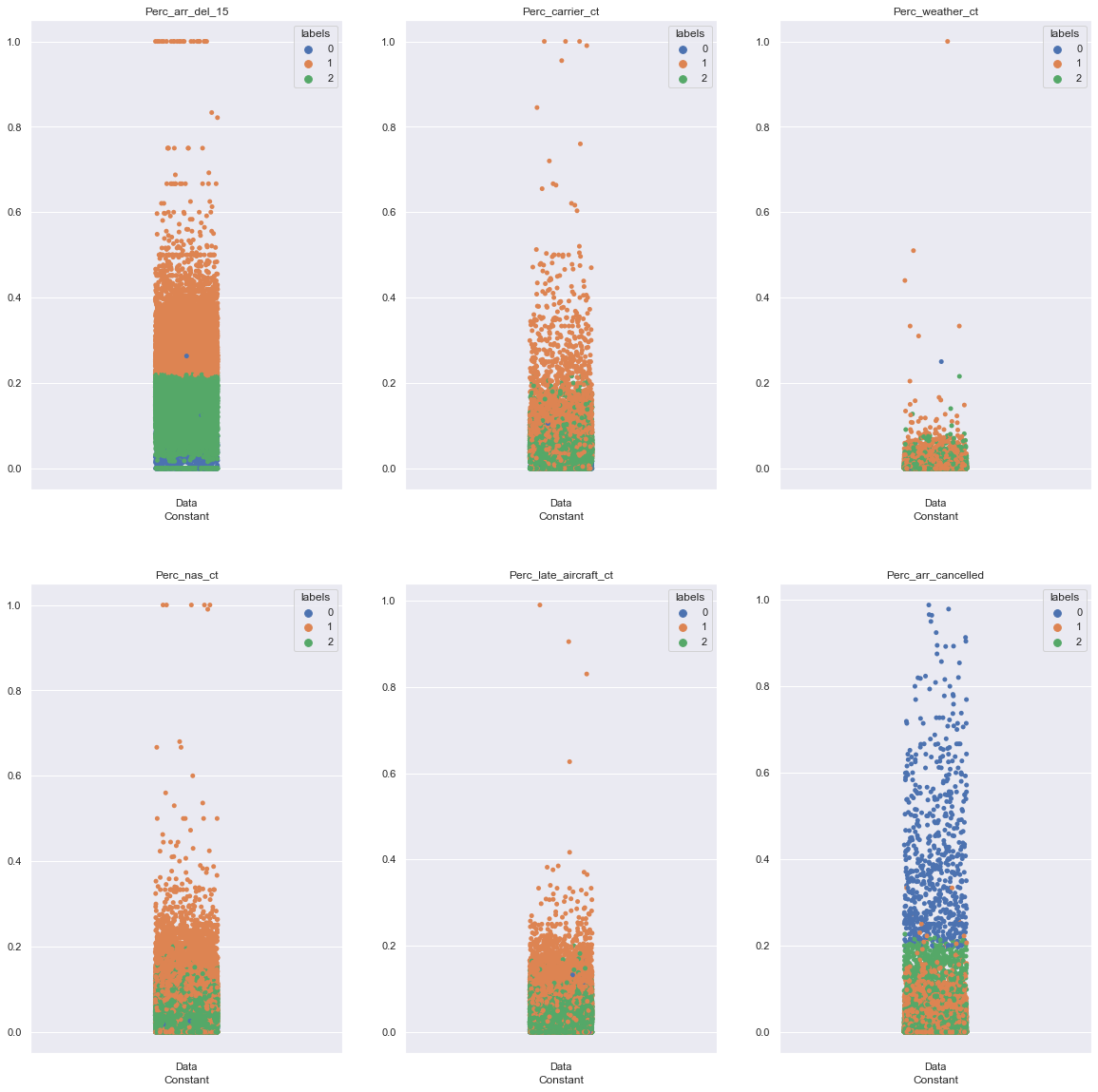
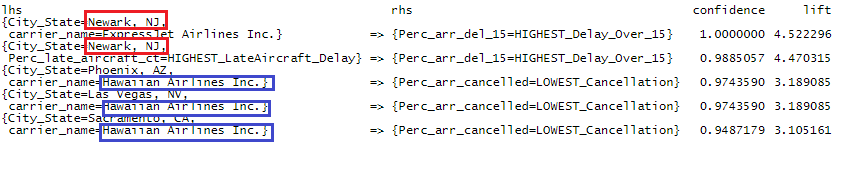


Figure 5. K-means clustering results of the flight data for two different dimensions

## Flight Data - Association rules mining results

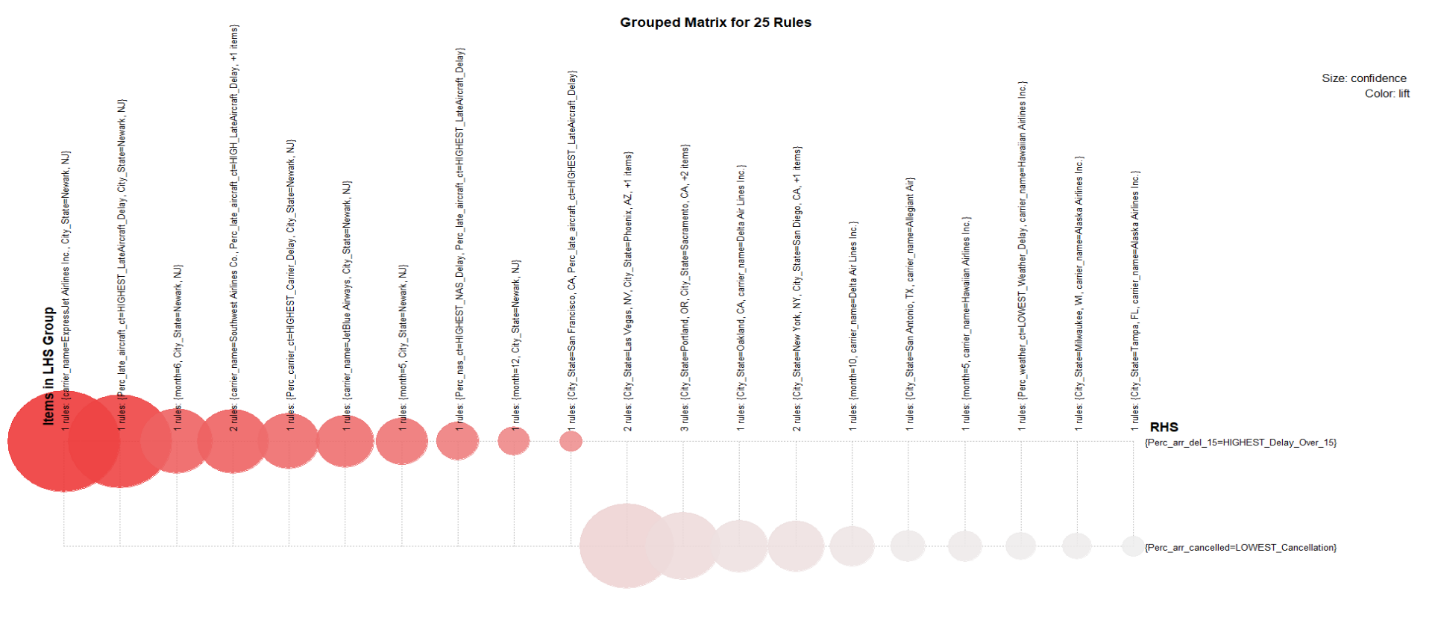
Due to the large variety of airports, carriers, and features, the support for most rules is low. The analysis focuses on rules with high confidence and lift scores, which respectively measure how frequently certain items appear with other items in an association and the ratio of the confidence of the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

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.

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

## Flight Data - Overall Results Analysis

Perhaps the most important feature 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 other words, avoid flights through this carrier in Newark if possible.

Overall, the way the data can be used to answer these questions demonstrate how data science results can identify patterns in data via visualization, statistical analysis, and data mining. These identified patterns can ultimately lead to the development of an action plan which lead to improvements in efficiency and operations for any given industry. Additionally, the data visualizations provide an efficient way to communicate the data and its analysis to relevant professionals. They highlight and display the key findings and conclusions to people who may not be privy to machine learning methods and data science.

# IST 718 - Big Data Analytics

In addition to classical machine learning methods, IST 718 introduces newer data science techniques such as neural networks including Pytorch and Tensorflow packages in Python 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 - 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. The model was created by utilizing Prophet, which is an additive regression model created by Facebook that is particularly strong at forecasting data with daily observations and strong seasonalities.

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 symptoms appear to the reporting of a Covid-19 related death is about 21 days.

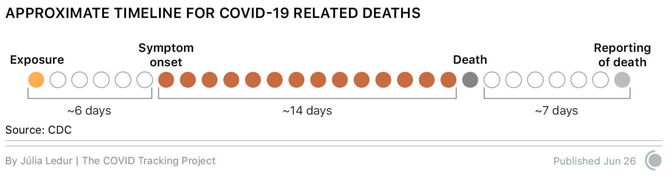


Figure 8: Timeline of Covid-19 related deaths

To account for this variable range in lag between cases and deaths, the Prophet Model will include 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 will predict the Covid-19 deaths for each state 21 days into the future.

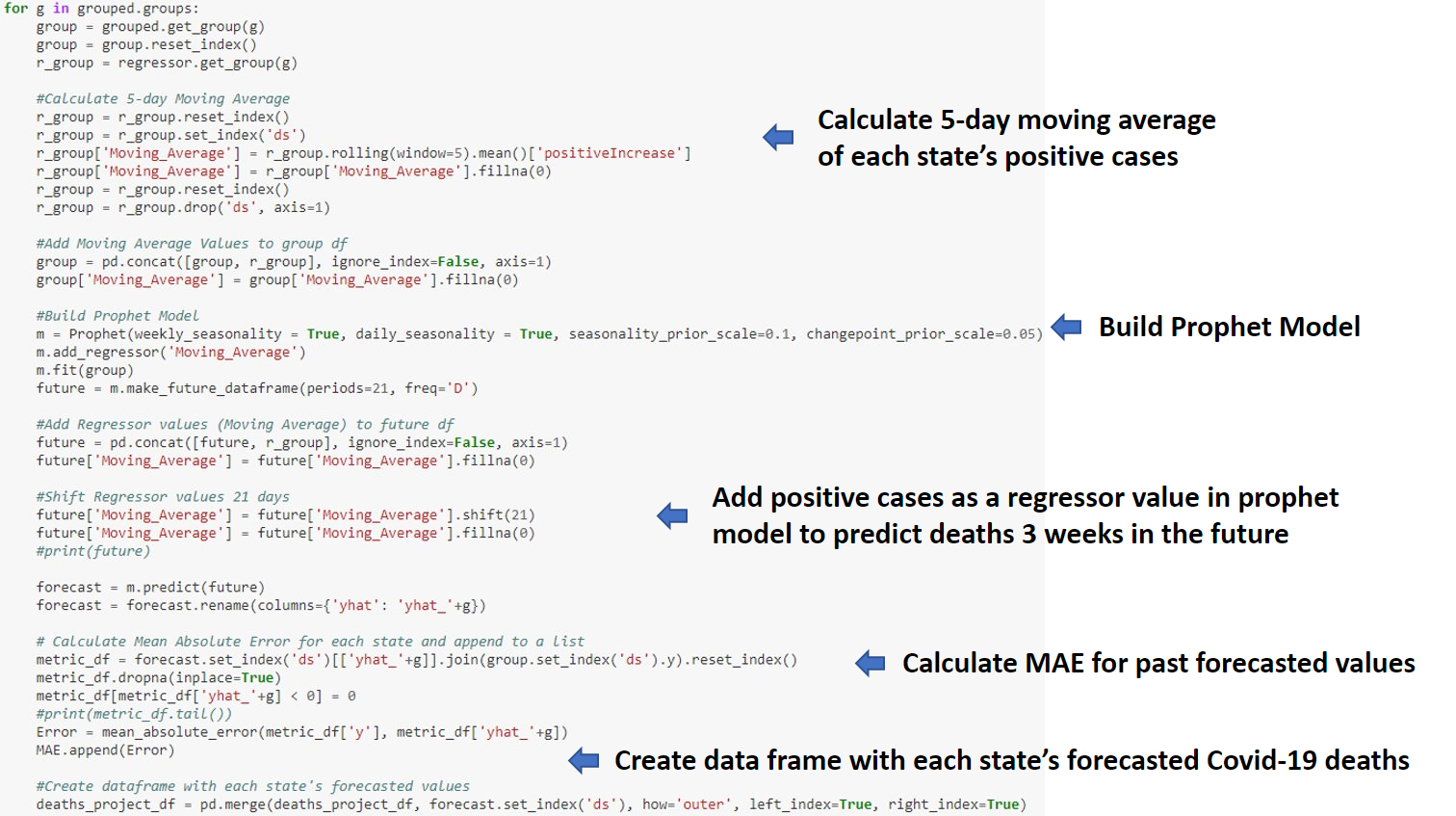


Figure 9: Components of Prophet Model in Python to predict Covid-19 related deaths

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.

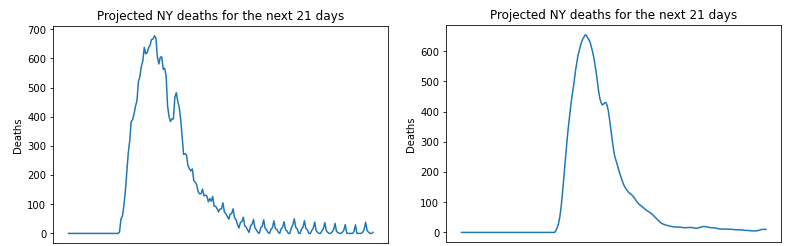


Figure 10. The left and right graph show the Prophet forecasting results for the state of NY before and after smoothing respectively

## 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](https://public.tableau.com/profile/joe.reyes3839#!/vizhome/ProphetForecastofCovid-19Deaths/Story1) via a Google Sheets API. This connection allows the dashboard to be refreshed as the Covid Tracking Project continues to publish daily numbers.

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.

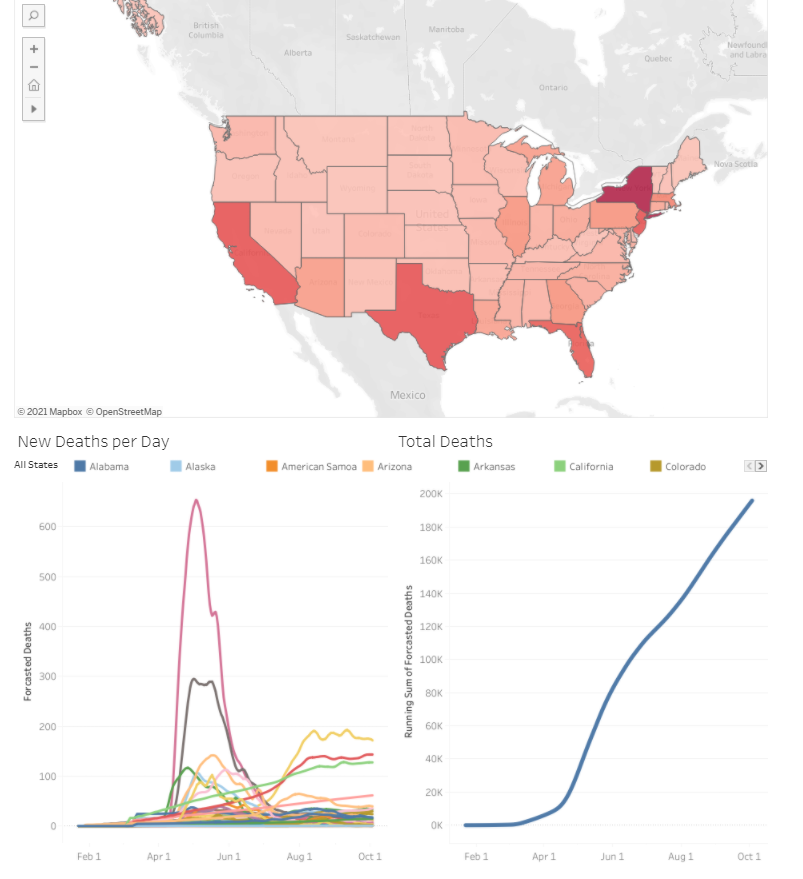


Figure 11. Overview of Prophet Results on Tableau. The map displays the Covid-19 related deaths for each state. The bottom-left line graph displays the time series forecasting for all states while the bottom-right lien graph shows the running tally of total Covid-19 related deaths in the United States.

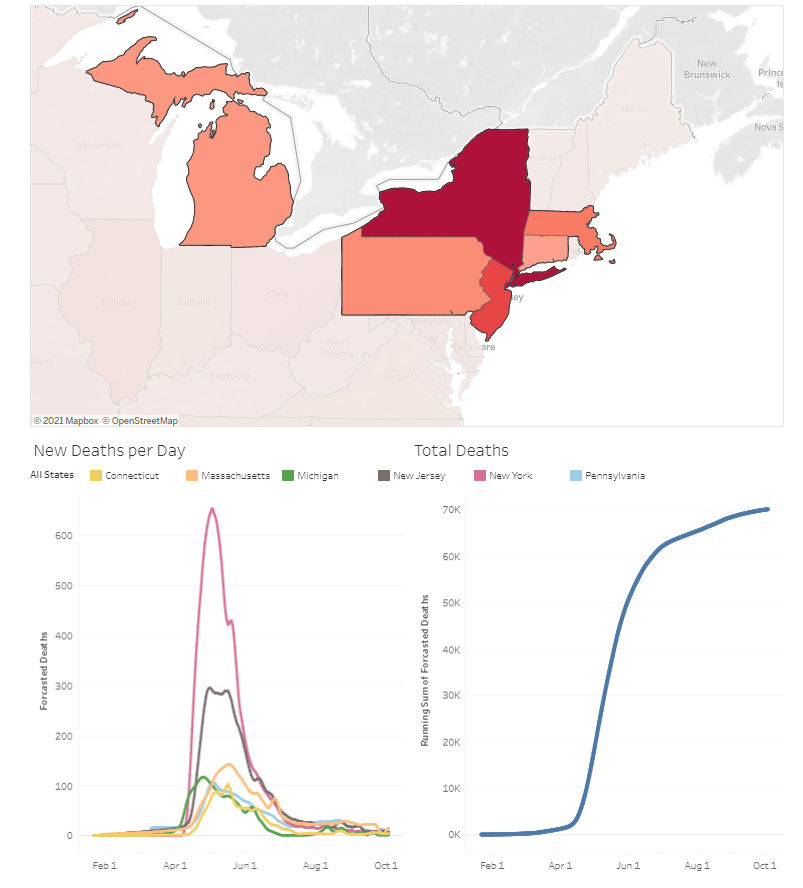


Figure 12. Prophet Results on Tableau of states in the Northeast and Midwest.

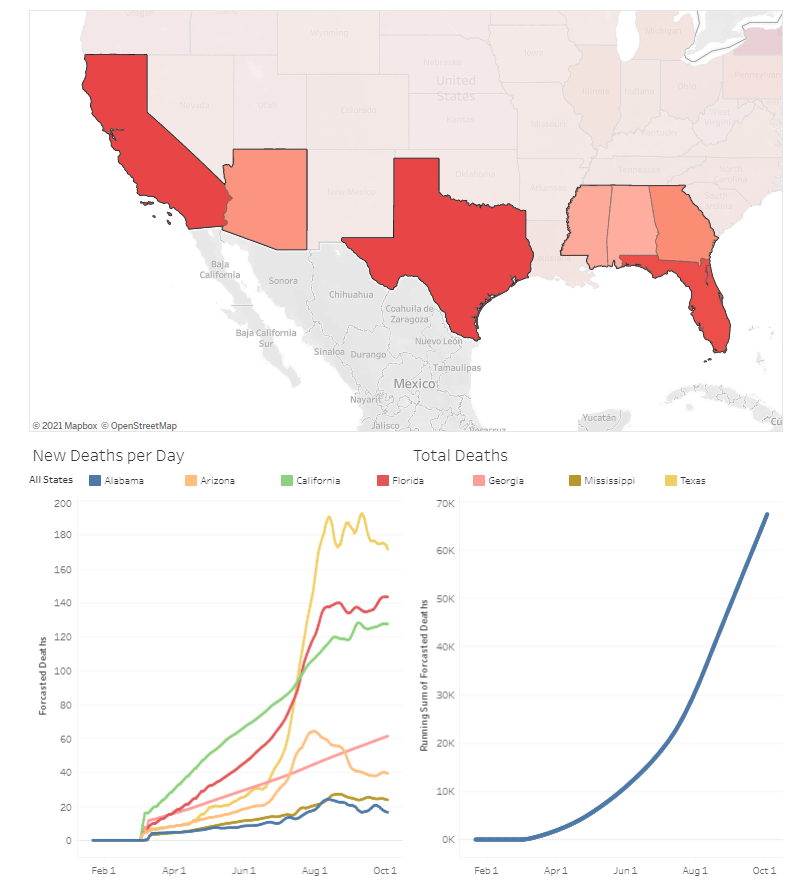


Figure 13. Prophet Results on Tableau of states in the Sun Belt.

## Covid-19 Deaths - Overall Results Analysis

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 in the time series forecasting provide what is perhaps the most advantageous feature of predictive analytics: using data to predict future outcomes and create a pre-emptive action plan to implement the business decisions derived from the analyses.

The data 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? Like the flight data results, the Covid-19 results show how data can be used to identify patterns and develop alternative strategies based on the data. 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?

Lastly, 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 will also continue to better predict deaths as it collects more and more historical data for use in its predictive analysis. Overall, this is a testament to the fact that data science is a dynamic discipline and will only continue to advance as more data is gathered and data mining techniques are improved.

# IST 652 – Scripting for Data Analysis

One of the 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 often produce superior results when compared to more traditional tasks such as support vector machine (SVM) or k-means clustering.

The use of deep learning is demonstrated in a Toxic Comment Classification project for IST 652. The data comes from a Kaggle Competition where contestants are challenged to create a model that predicts a text’s category of toxicity as either toxic, severely toxic, obscene, threatening, insulting, or identity hate.

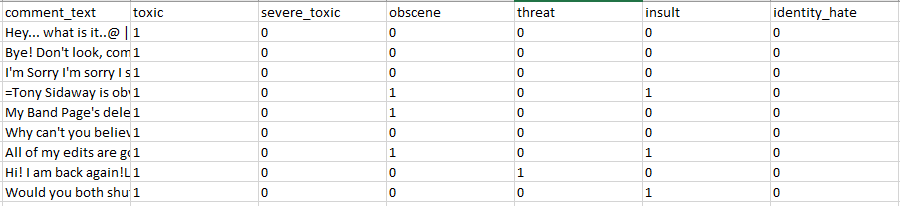


Figure 14. Sample of data from the Toxic Comment Classification Challenge. Each comment has been scored by a human based on its toxicity.

## Toxic Comments – Neural Networks

TenserFlow 2.0, an open-source machine learning platform, was utilized to create a long short term memory model (LSTM). LSTMs are 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 used 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.

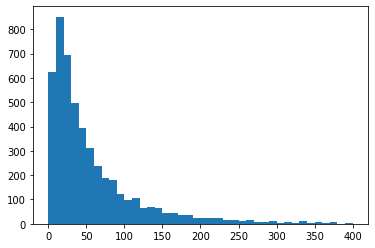


Figure 15. Distribution of word lengths in the Toxic Comments dataset

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, also known as 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.

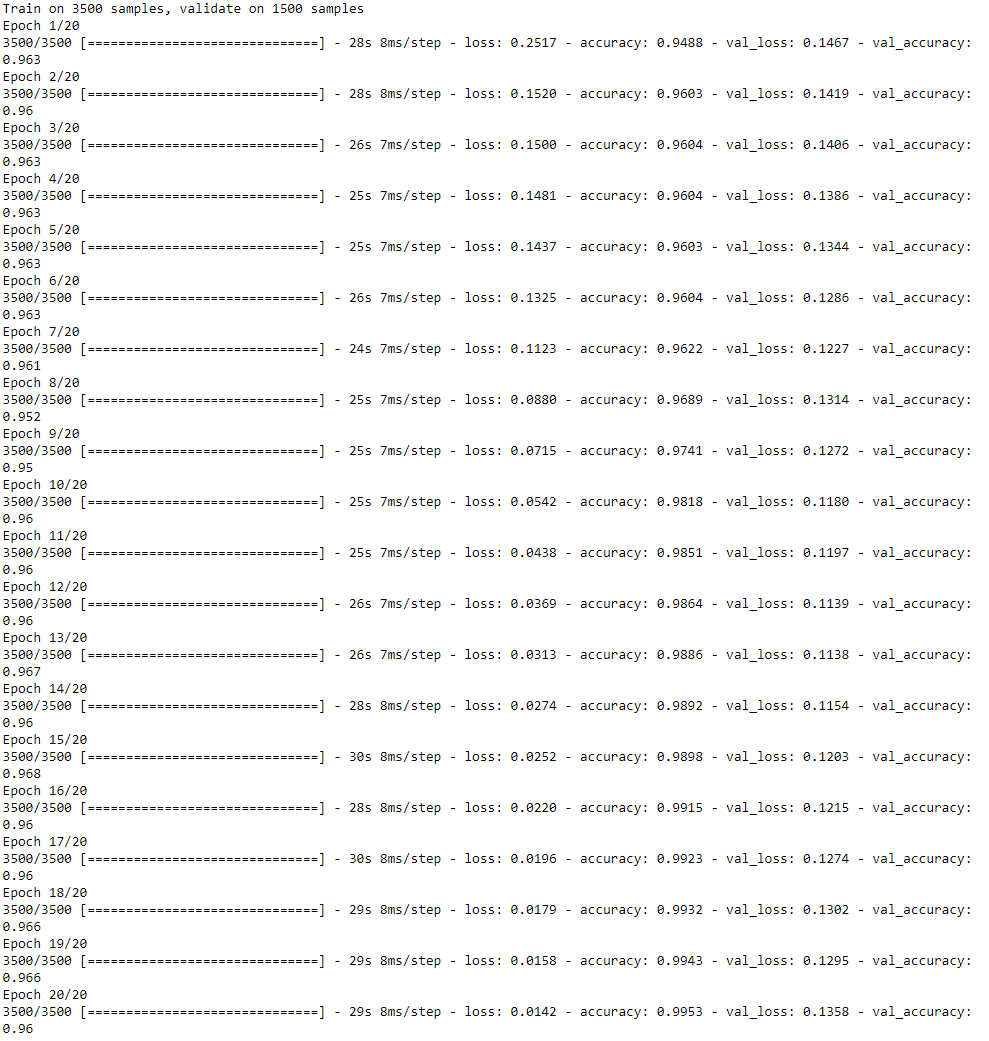


Figure 17. LSTM Model Results for Toxic Comments Classification

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 - Overall Results Analysis

In essence, 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 perhaps one of the most discussed controversial topics in data science. While there is certainly a need to protect patients and their health information, certain fields are helpful and sometimes even necessary for use in medical research and clinical trials. The 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 will be designed for a clinical data registry, which collects and stores study data. This database will focus 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) and date of birth (DOB) and patient zip code are considered PHI and must be de-identified.

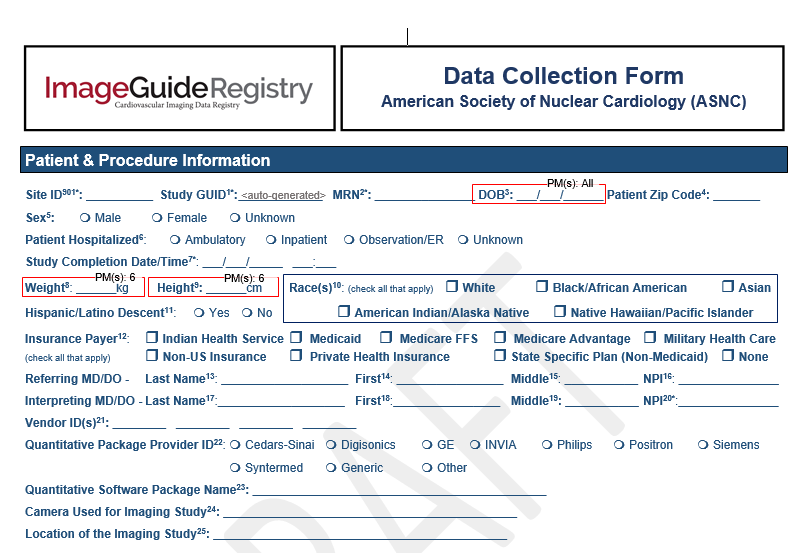


Figure 18. Sample of medical report form that is collected in the ImageGuide Registry

To focus on areas that can be used for research and quality improvement, the database will extract 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.

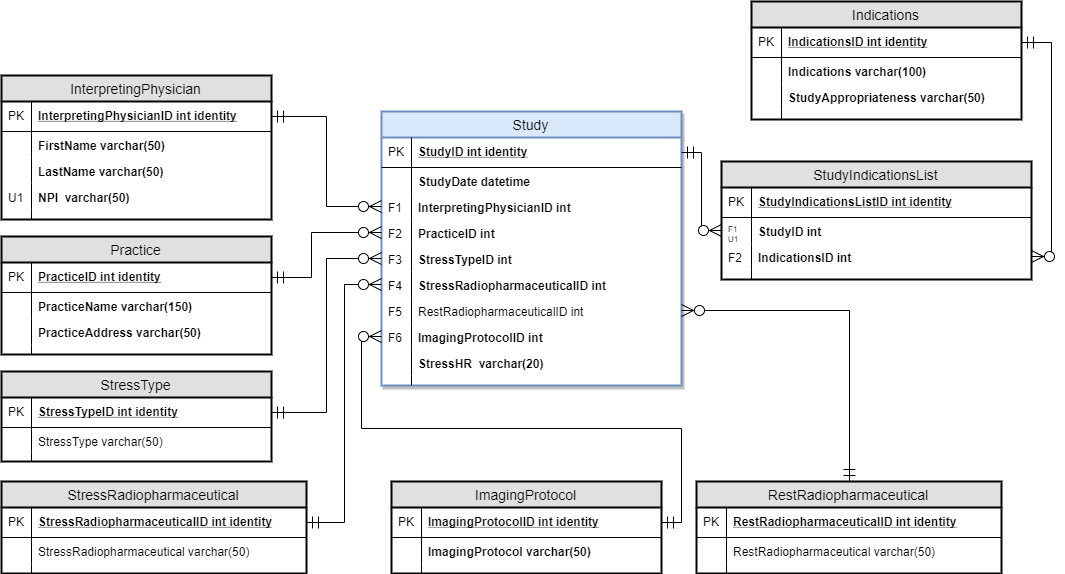
****

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.

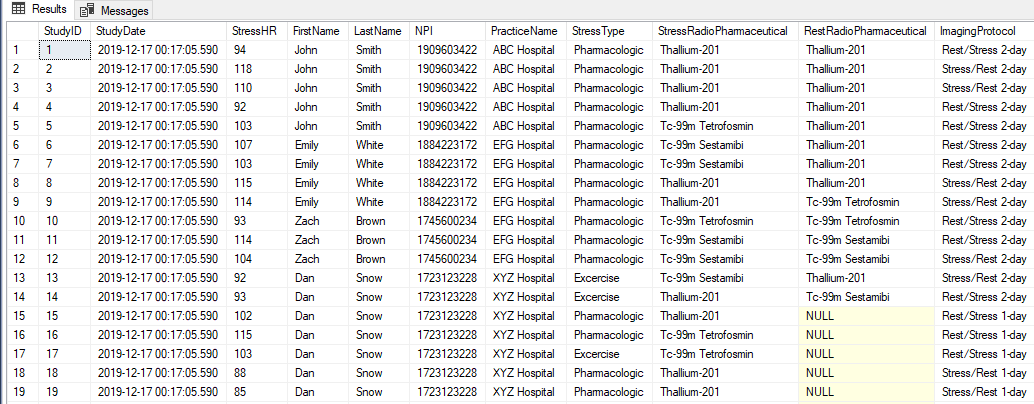


Figure 21. Sample of SQL database with de-identified medical record data

## Nuclear Cardiology Database – Overall Analysis

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.

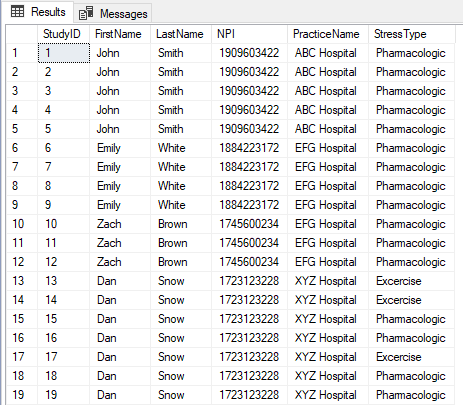
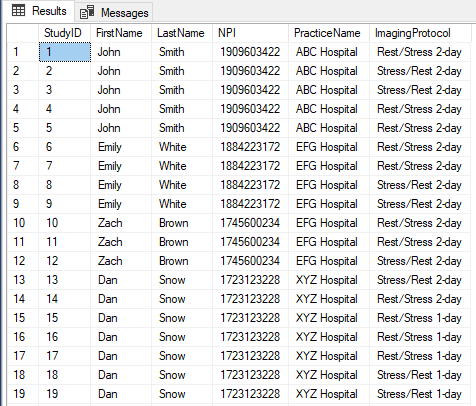
 

Figure 22. Sample of SQL query results that answer specific data questions

This data is 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.

Lastly, it is important to note that the database is strictly de-identified and lacks any 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 through 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 overall 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, data science is a lifelong learning mission and this is only just the beginning.