The Applied Data Science program at Syracuse University's School of Information Studies provides students the opportunity to collect, manage, analyze, and develop insights using data from a multitude of domains using various tools and techniques. In courses such as Database Administration (Peoples, "IST 659,", 2018), Data Analytics (Peoples, "IST 707,", 2018), Text Mining (Peoples, "IST 736,", 2018), and Marketing Analytics (Peoples, "MAR 653,", 2018), reports and presentations were developed to deliver insights using Microsoft Access, SQL Server Management Studio, Python, R, Excel and Tableau.

Through studying Database Administration under the direction of Dr. Gregory Block, a Music Database was developed to organize music artists, songs, and artist metadata including genre, instrument, recording studio, awards, and personal play counts (Peoples, "IST 659,", 2018). In development and population of the database, the scope of the implementation was reduced to three albums for each of five artists, each associated with at least one award; this application required the collection and organization of data to develop actionable insights, as well as the development of alternative strategies based on the domain, scope, and context of the problem.

Conceptual and logical models (Fig. 1) were developed to organize the relationships between band members, instruments, recording studios, artists, albums, awards, and songs. Tables were created in SQL Server Management Studio while data population was accomplished using Microsoft Access, which also facilitated the exploration of data. Reports and stored procedures were created to display the earliest and latest album release date, the most popular genre, the album and artist with the most awards, the most popular artist hometown, and the most played songs in the database (Fig. 2). These questions provide valuable insights into the aspects which contribute to a user's music interests, which can be leveraged to discover similar artists.

Logical Model

Fig. 1: Logical Model, (Peoples, "IST 659,", 2018).

Most Played Songs

Thursday, September 13, 2018
4:13:28 PM

song_name	num_plays	artist_name	album_name
Love, Reign O'er Me	1988	The Who	Quadrophenia
The Luckiest	1940	Ben Folds	Rockin' the Suburbs
Landslide	1884	Fleetwood Mac	The White Album
Its Not True	1876	The Who	My Generation
Babe I'm Gonna Leave You	1864	Led Zeppelin	Led Zeppelin
The Chain	1844	Fleetwood Mac	Rumours
Whole Lotta Love	1664	Led Zeppelin	Led Zeppelin II
Out in the Street	1651	The Who	My Generation
Dani California	1646	Red Hot Chili Peppers	Stadium Arcadium
Scar Tissue	1646	Red Hot Chili Peppers	Californication

Fig. 2: Most Played Songs, (Peoples, "IST 659,", 2018).

The exercise of developing a data management solution revealed the significance of how the data is stored and accessed, which is imperative to analysts and data scientists. In the final term

DRAFT

of the Applied Data Science program, courses in Advanced Database Management and Data Warehousing will explore more complex database architectures, security, and performance tuning to make more informed and efficient business decisions. This project also contributed to the ability to deliver actionable insights in the field of marketing analytics; which often requires precise and efficient information management to account for customer promotion participation and measurement of the performance of marketing campaigns. Automation of operational tasks such as direct mail programs and promotional tagging is easily accomplished with the skills developed in the Applied Data Science program and provide value to marketing analytics teams.

Through studying Data Analytics under the direction of Dr. Mohammed Syed, various data mining techniques were introduced which perform with varying precision and efficiency for applications in regression, classification, and clustering. In the final presentation, Naïve Bayes and Decision Tree Classification techniques were implemented to compare computation time and accuracy in predicting the edibility of mushrooms (Peoples, "IST 707,", 2018). R Studio is leveraged to conduct analysis within r using data from *Kaggle*, a public data mining resource.

This application required the cleaning and preprocessing of data, which involved the transformation of twenty-two categorical features into ninety-six binary features for use in classification. This also facilitated identifying patterns in the data, such as the most predictive features for poisonous mushrooms, the most ubiquitous features among both edible and poisonous mushrooms, and the certainty by which edibility can be predicted. In comparisons made by Jin Huang, accuracy between Decision Trees and Naïve Bayes were comparable, with preference on the latter for ranked classification (Huang, Lu, & Ling). Similarly, research submitted by Pranckevičius & Marcinkevičius found that decision trees perform with lower accuracy when compared to Naïve Bayes classifiers (Pranckevičius & Marcinkevičius, 2017); recent findings

suggest Naïve Bayes is the preferred classification algorithm over decision trees. However, observations made in this application revealed nearly three percent greater accuracy using decision trees in one-tenth of the time when compared to Naïve Bayes; decision trees predicted the edibility of mushrooms with 99.14% accuracy in 1.7 seconds of training time (Fig. 3).

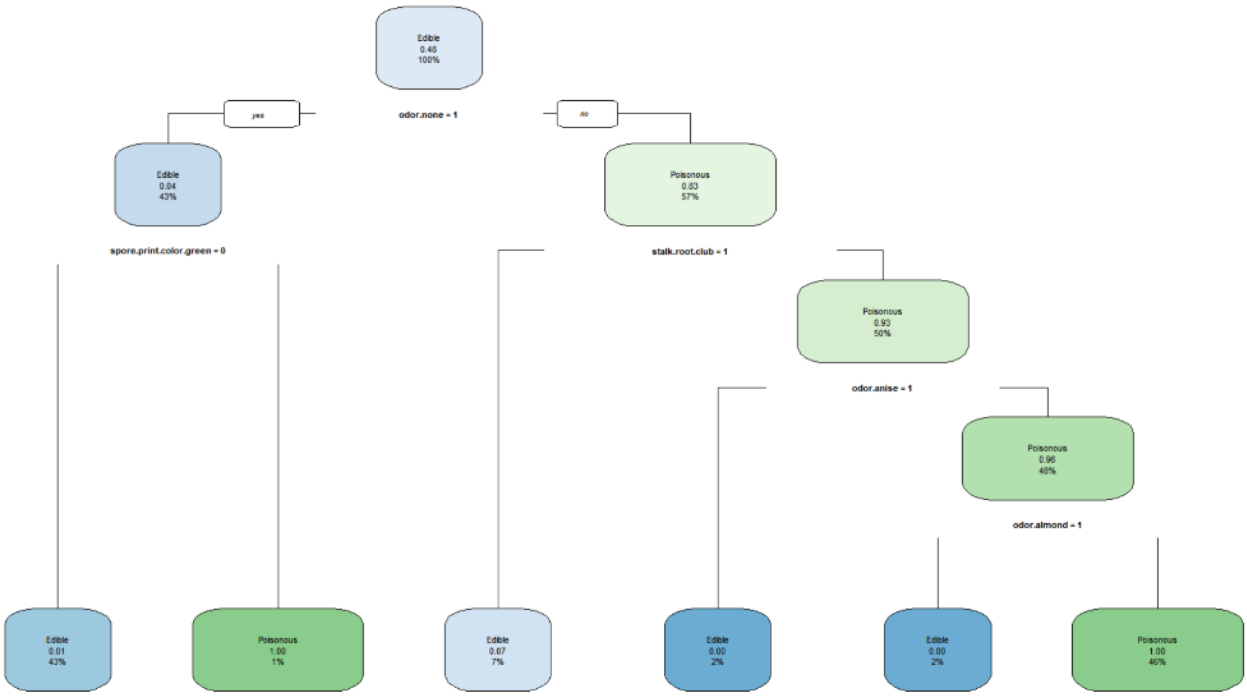

Fig 3. Decision Tree Model, (Peoples, "IST 707,", 2018).

This unexpected result is an example of the importance of testing different data mining techniques to develop the simplest, most accurate prediction models. Testing alternative strategies and weighing the benefits of each technique with respect to the data can reduce computation costs and provide the greatest precision in data mining tasks. In a marketing analytics setting, this is an important distinction that is magnified by the scale of the data. With more heavy computation being done using services such as Azure, the consideration of computational costs is growing in significance.

Through studying Text Mining under the direction of Dr. Bei Yu, data mining techniques were introduced to analyze text and develop insights from unstructured data. In the final

presentation, Latent Dirichlet Allocation is leveraged in tandem with K-Means Clustering to group Congressmen by the content of their Tweets, thereby providing insight into the policies by which they stand. This technique was also successfully implemented using unstructured data by Akash Bhoi which classified healthy and unhealthy gaits in Parkinson's patients (Bhoi, 2017). Miha Pavlinek also found success using LDA to improve classification tasks using *Mallet* (Pavlinek & Podgorelec, 2017). In this exercise, *Python* is used for data collection and clustering while *Mallet* is selected to generate topic models; exploratory data analysis is conducted in *Tableau*.

This application required the collection of 527 Twitter usernames belonging to members of the United States Congress, as well as the latest two hundred Tweets from each account. Once all Tweets had been collected, each member's corpus was saved to a text file, and topic models were created using *Mallet*. Fifty clusters are selected to capture a manageable number of topics, with testing and inspection of ten, fifteen, and twenty-five clusters. These weights are subsequently paired with each member's party affiliation, house seat, and seniority in years before being clustered; the elbow-method is used to observe an optimal number of clusters of $k=3$ (Fig. 4).

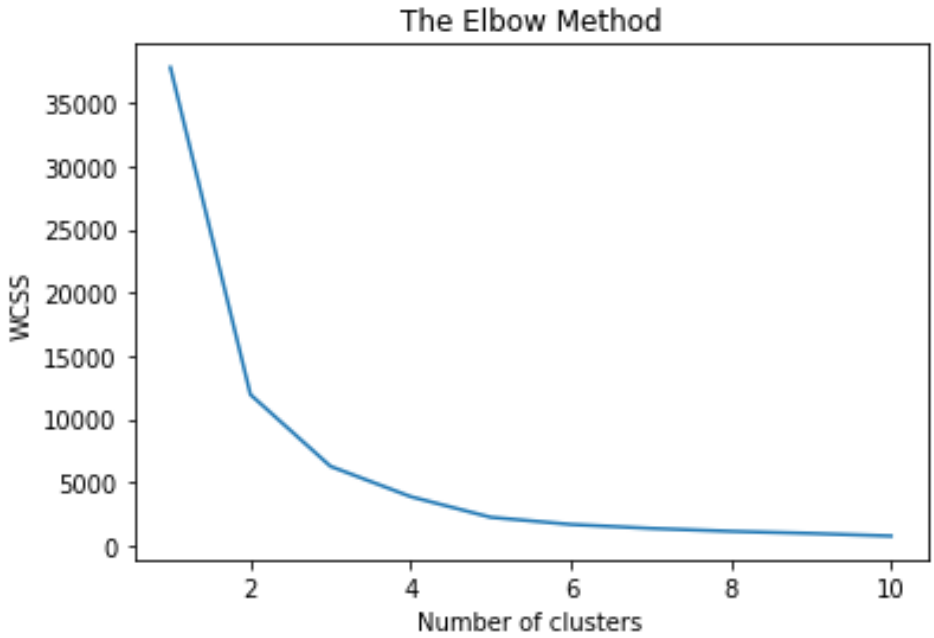

Fig 4. Elbow Method, (Peoples, "IST 736,", 2019).

Following clustering, the data is exported to *Tableau* and profiled based on seniority and party affiliation (Fig. 5). Three distinct groups are observed, where those with less than five years of experience are typically identified as Freshmen, members with an average of roughly twenty-five years of experience are clustered as Old Guard, and members that participate in more rhetoric-filled debate and have ranging seniority are clustered as Media Players. Attribution is then made to the labeled clusters by averaging the LDA weights across each cluster and subtracting that from the average of the weights within the cluster (Fig. 6). This provides the magnitude of weight that is attributed to each topic. Universal topics are identified by weight less than 0.001, and biased group-specific topics have weights greater than 0.0025. Ten universal topics are observed, and sixteen are labeled as group-specific.

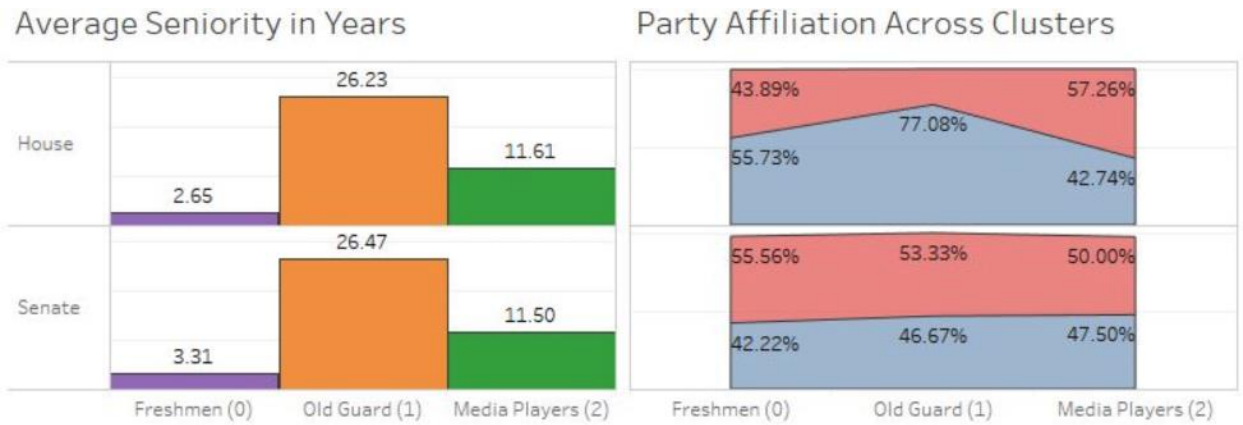

Fig. 5. Segmentation Profiling, (Peoples, "IST 736,", 2019).

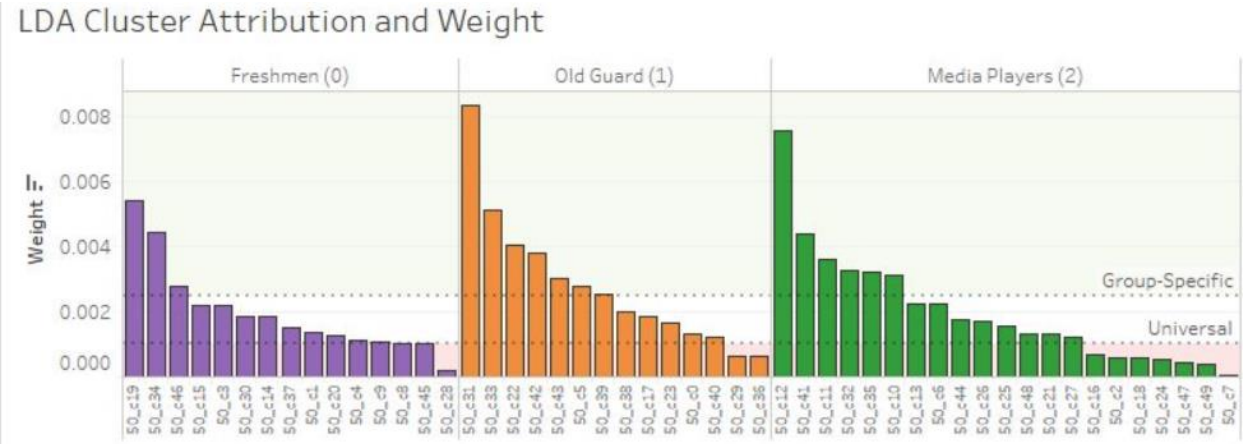

Fig. 5. Topic Cluster Attribution, (Peoples, "IST 736,", 2019).

DRAFT

This exercise provided the opportunity for the collection and structuring of externally sourced data, identification of patterns within and between clusters of text and developed insights into the behavior of elected officials. The collection of data required the consideration of user privacy, where only the Tweets made by Congressional members were collected, rather than capturing unnecessary information. Although all the collected data is public record, considerations must be made to ensure that only the relevant information is requested to both balance request limitations and user privacy. In the final term of the Applied Data Science program, additional study of user privacy and information management will take place in Information Policy, where topics encompassing intellectual property and freedom of expression will be explored.

Text data is incredibly important to marketing analytics teams as more unstructured sources are introduced. With more content being created by customers such as reviews, social media posts, and transcriptions, the value of extracting quantifiable and actionable insights from text is growing in significance. As organizations project a greater social media presence, the ability to organize and analyze large collections of text allows for automation using conversational assistants, as well as predictive analytics with text mining.

Through studying Marketing Analytics under the direction of Dr. Andrew Petersen, data mining concepts specific to marketing were introduced which inspired the final presentation where yearly transaction information for households from a grocery store was used to identify a target group of customers using K-Means Clustering, and subsequently derive an optimal promotional offer using Apriori Rule Association and Sensitivity Analysis for use in a direct mail marketing campaign. The tools required included *Python* for clustering and rule association, while *Excel* was used for exploratory data analysis.

This exercise involved exploring and cleaning the data prior to clustering, by transforming categorical columns into one-hot encoded vectors. Segmentation was completed using the items purchased and pricing data, while profiling was accomplished using demographic information such as income, age, and household size; considerations are made to not cluster customers on demographic data to avoid biases being introduced to the models. The selected cluster contained nearly twenty percent of the entire customer base and use coupons slightly more often than the average customer. Three optimal carts are selected using Apriori Rule Association on the items purchased by customers within the selected group, with coupon promotions applied to two items (Fig. 7). These items are selected such that when items from *cart one* are purchased, items from *cart two* have a discount applied. This method was also successfully implemented by Grazyna Suchacka in predicting purchase behavior in an e-commerce setting; this study similarly selected a target subset of their customers to improve the expected result (Suchacka & Chodak, 2016). The optimal discount was then calculated using the average discounts applied previously and varying the percent discount to not exceed a $10,000 liability margin (Fig. 8). This resulted in an expected 137% increase in gross revenue for the affected products.

Cluster	Cart1 ->	Cart2	conf	supp	lift	conv
2	drinks, frozen_pizza	meat	0.943	0.085	1.265	4.454
2	baking, food	food_add-ons	0.928	0.217	1.417	4.803
2	dessert, packaged_foods	meat	0.928	0.083	1.244	3.512

Fig. 7. Optimal Apriori Rules, (Peoples, "MAR 653,", 2019).

Expected participants:	8500
Maximum liability:	$10,000
Average price of meat:	$3.58
○ Average coupon value:	$0.26
○ Discount:	**7.3%**
○ Num offers:	2
Average price of food add-on:	$1.97
○ Average coupon value:	$0.28
○ Discount:	**14.0%**
○ Num offers:	1

Liability = 8500
*((2*3.58**MeatDiscount*)
+(1*1.97**FoodAdd-OnDiscount*))

meat/food	0	0.01	0.02	0.03	0.04	0.05	0.06	0.07
0	0	166.6	333.2	499.8	666.4	833	999.6	1166.2
0.01	608.6	775.2	941.8	1108.4	1275	1441.6	1608.2	1774.8
0.02	1217.2	1383.8	1550.4	1717	1883.6	2050.2	2216.8	2383.4
0.03	1825.8	1992.4	2159	2325.6	2492.2	2658.8	2825.4	2992
0.04	2434.4	2601	2767.6	2934.2	3100.8	3267.4	3434	3600.6
0.05	3043	3209.6	3376.2	3542.8	3709.4	3876	4042.6	4209.2
0.06	3651.6	3818.2	3984.8	4151.4	4318	4484.6	4651.2	4817.8
0.07	4260.2	4426.8	4593.4	4760	4926.6	5093.2	5259.8	5426.4
0.08	4868.8	5035.4	5202	5368.6	5535.2	5701.8	5868.4	6035
0.09	5477.4	5644	5810.6	5977.2	6143.8	6310.4	6477	6643.6
0.1	6086	6252.6	6419.2	6585.8	6752.4	6919	7085.6	7252.2
0.11	6694.6	6861.2	7027.8	7194.4	7361	7527.6	7694.2	7860.8
0.12	7303.2	7469.8	7636.4	7803	7969.6	8136.2	8302.8	8469.4
0.13	7911.8	8078.4	8245	8411.6	8578.2	8744.8	8911.4	9078
0.14	8520.4	8687	8853.6	9020.2	9186.8	9353.4	9520	9686.6

Fig. 8. Calculation of Optimal Coupon Value, (Peoples, "MAR 653,", 2019)

DRAFT

This project provided the opportunity to organize and analyze transaction information using data mining techniques, as well as visualization to identify patterns for customer targeting. It was also necessary to develop a plan of action to quantify the insights developed in this analysis, which translates to measurable and actionable recommendations. Ethical considerations were also necessary to ensure that customer segmentation and profiling was free of bias, using demographic information to profile the previous behavior of a customer, rather than using said information to explain their behavior. This project allowed the data to guide the analysis, requiring alternative strategies to be developed as observations were made within the data.

The Applied Data Science program at Syracuse University's School of Information Studies provides students the opportunity to synthesize the collection, management, analysis of data, and delivery of actionable insights using various data science techniques. Skills learned in the program have developed a multifaceted approach to solving structured and unstructured data problems, it has also cultivated strategies that improve organizational efficiency. The program has fostered a practice of transparency, reproducibility, and ethical data management which promotes integrity and credibility within an organization's analytics team. Using the methods learned at the School of Information Studies, data scientists are equipped with the ability to tackle a wide range of problems and the resources to explain observations to a variety of stakeholders and business professionals.

References

Bhoi, A. K. (2017). Classification and Clustering of Parkinson's and Healthy Control Gait Dynamics Using LDA and K-means. *Int. J. Bio Automation*, 21(1), 19-30. Retrieved from http://www.biomed.bas.bg/bioautomation/2017/vol_21.1/files/21.1_02.pdf

Huang, J., Lu, J., & Ling, C. (n.d.). Comparing naive Bayes, decision trees, and SVM with AUC and accuracy. *Third IEEE International Conference on Data Mining*. doi:10.1109/icdm.2003.1250975

Pavlinek, M., & Podgorelec, V. (2017). Text classification method based on self-training and LDA topic models. *Expert Systems with Applications*, 80, 83-93. doi:10.1016/j.eswa.2017.03.020

Peoples, S. L. (n.d.). (2018) IST 659: Database Administration. Retrieved from https://github.com/SLPeoples/MSADS_Portfolio/tree/master /IST659_DatabaseAdministration

Peoples, S. L. (n.d.). (2018) IST 707: Data Analytics. Retrieved from https://github.com/SLPeoples/MSADS_Portfolio/tree/master/IST707_DataAnalytics

Peoples, S. L. (n.d.). (2019) IST 736: Text Mining. Retrieved from https://github.com/SLPeoples/MSADS_Portfolio/tree/master/IST736_TextMining

Peoples, S. L. (n.d.). MAR 653: Marketing Analytics. Retrieved from https://github.com/SLPeoples/MSADS_Portfolio/tree/master /MAR653_Marketing_Analytics

Pranckevičius, T., & Marcinkevičius, V. (2017). Comparison of Naive Bayes, Random Forest, Decision Tree, Support Vector Machines, and Logistic Regression Classifiers for Text Reviews Classification. *Baltic Journal of Modern Computing*,*5*(2). doi:10.22364/bjmc.2017.5.2.05

Suchacka, G., & Chodak, G. (2016). Using association rules to assess purchase probability in online stores. *Information Systems and E-Business Management*, 15(3), 751-780. doi:10.1007/s10257-016-0329-4