Syracuse University
School of Information Studies

Master of Science
Applied Data Science

Portfolio Milestone

Samuel L. Peoples

793568460
10 April 2019

https://github.com/SLPeoples/MSADS_Portfolio

Introduction

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.

Reports and presentations were created in courses which exemplify the skills developed in the program including, but not limited to:

- IST 659: Database Administration
- IST 707: Data Analytics
- IST 736: Text Mining
- MAR 653: Marketing Analytics

IST 659
Database Administration:

Music Database 4

Microsoft® SQL Server®

A

Introduction

- Through studying Database Administration under the direction of Dr. Gregory Block, a Music database was developed to organize the following data from a music collection:
  - Artists
  - Songs
  - Genre
  - Instrument
  - Recording Studio
  - Awards
  - Personal Play Counts

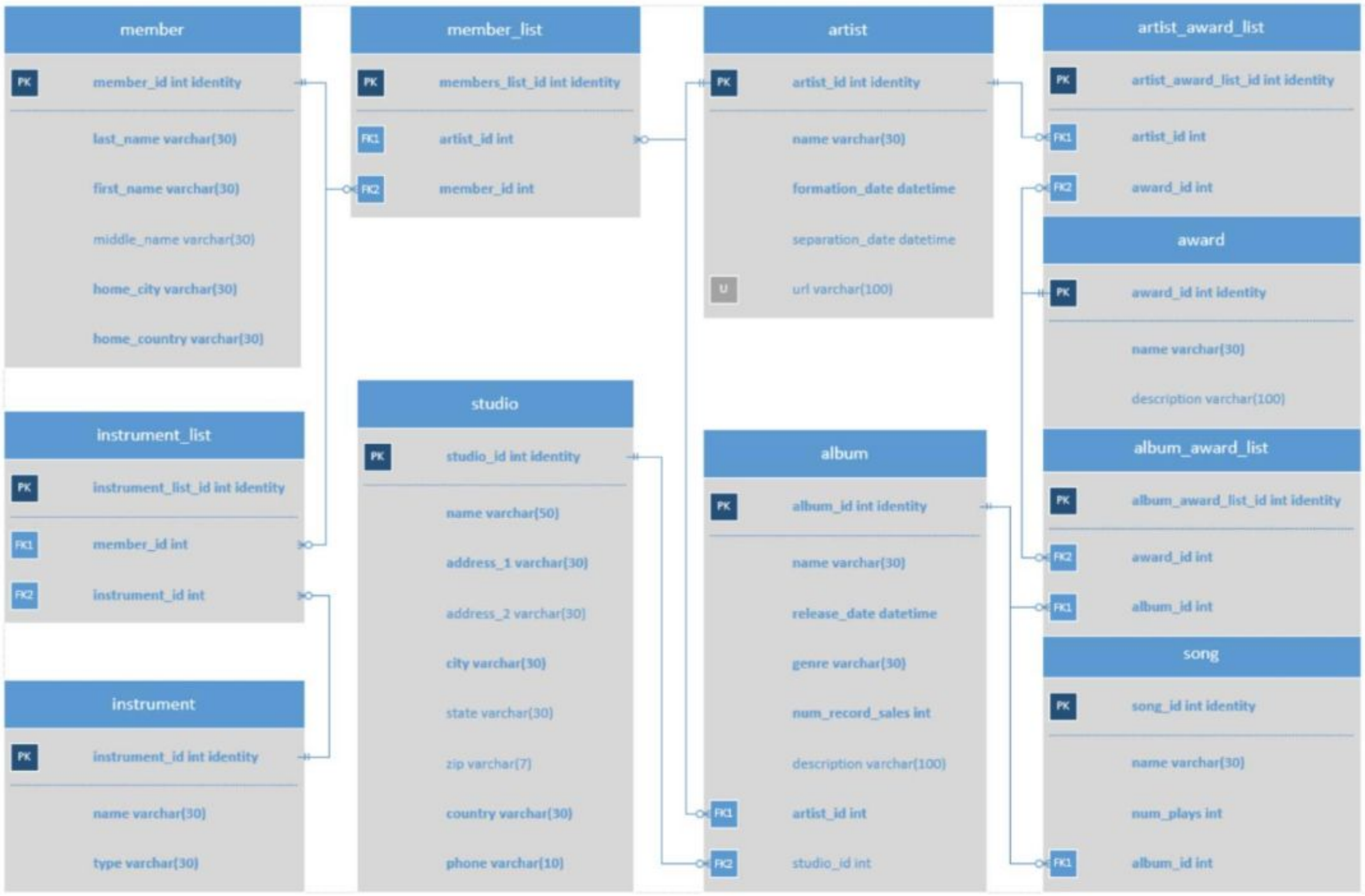
IST 659: Database Administration

Modeling, Table Creation, and Reporting

- Conceptual and Logical models were developed to organize the relationships between the various members of the database.
- SSMS was utilized to create and initially populate the tables.
- Stored procedures were created to report various summary statistics.
- Microsoft Access was used to facilitate bulk data entry and reporting.

IST 659: Database Administration

Logical Model



member
- PK member_id int identity
- last_name varchar(30)
- first_name varchar(30)
- middle_name varchar(30)
- home_city varchar(30)
- home_country varchar(30)

member_list
- PK members_list_id int identity
- FK1 artist_id int
- FK2 member_id int

artist
- PK artist_id int identity
- name varchar(30)
- formation_date datetime
- separation_date datetime
- U url varchar(100)

artist_award_list
- PK artist_award_list_id int identity
- FK1 artist_id int
- FK2 award_id int

award
- PK award_id int identity
- name varchar(30)
- description varchar(100)

instrument_list
- PK instrument_list_id int identity
- FK1 member_id int
- FK2 instrument_id int

studio
- PK studio_id int identity
- name varchar(50)
- address_1 varchar(30)
- address_2 varchar(30)
- city varchar(30)
- state varchar(30)
- zip varchar(7)
- country varchar(30)
- phone varchar(10)

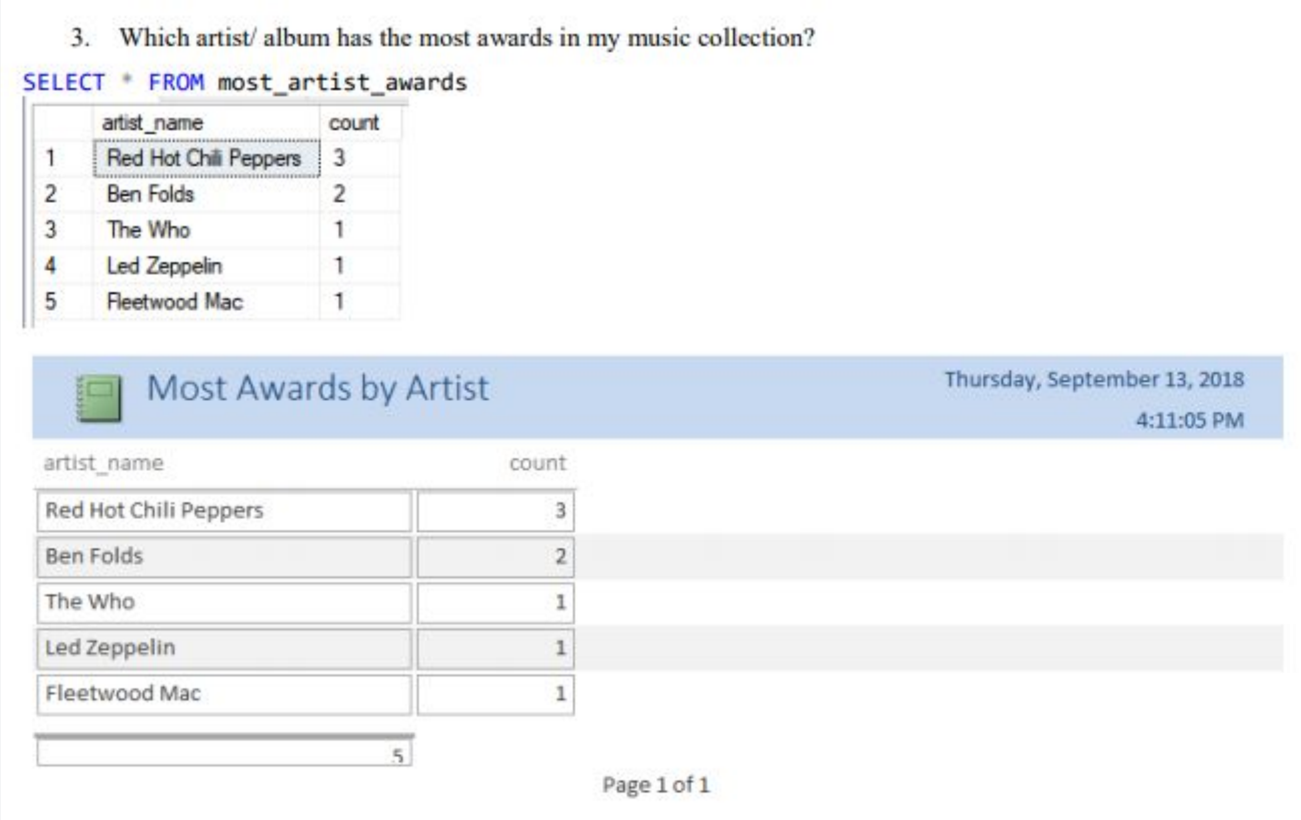
album
- PK album_id int identity
- name varchar(30)
- release_date datetime
- genre varchar(30)
- num_record_sales int
- description varchar(100)
- FK1 artist_id int
- FK2 studio_id int

album_award_list
- PK album_award_list_id int identity
- FK2 award_id int
- FK1 album_id int

instrument
- PK instrument_id int identity
- name varchar(30)
- type varchar(30)

song
- PK song_id int identity
- name varchar(30)
- num_plays int
- FK1 album_id int

IST 659: Database Administration



3. Which artist/ album has the most awards in my music collection?

```
SELECT * FROM most_artist_awards
```

	artist_name	count
1	Red Hot Chili Peppers	3
2	Ben Folds	2
3	The Who	1
4	Led Zeppelin	1
5	Fleetwood Mac	1

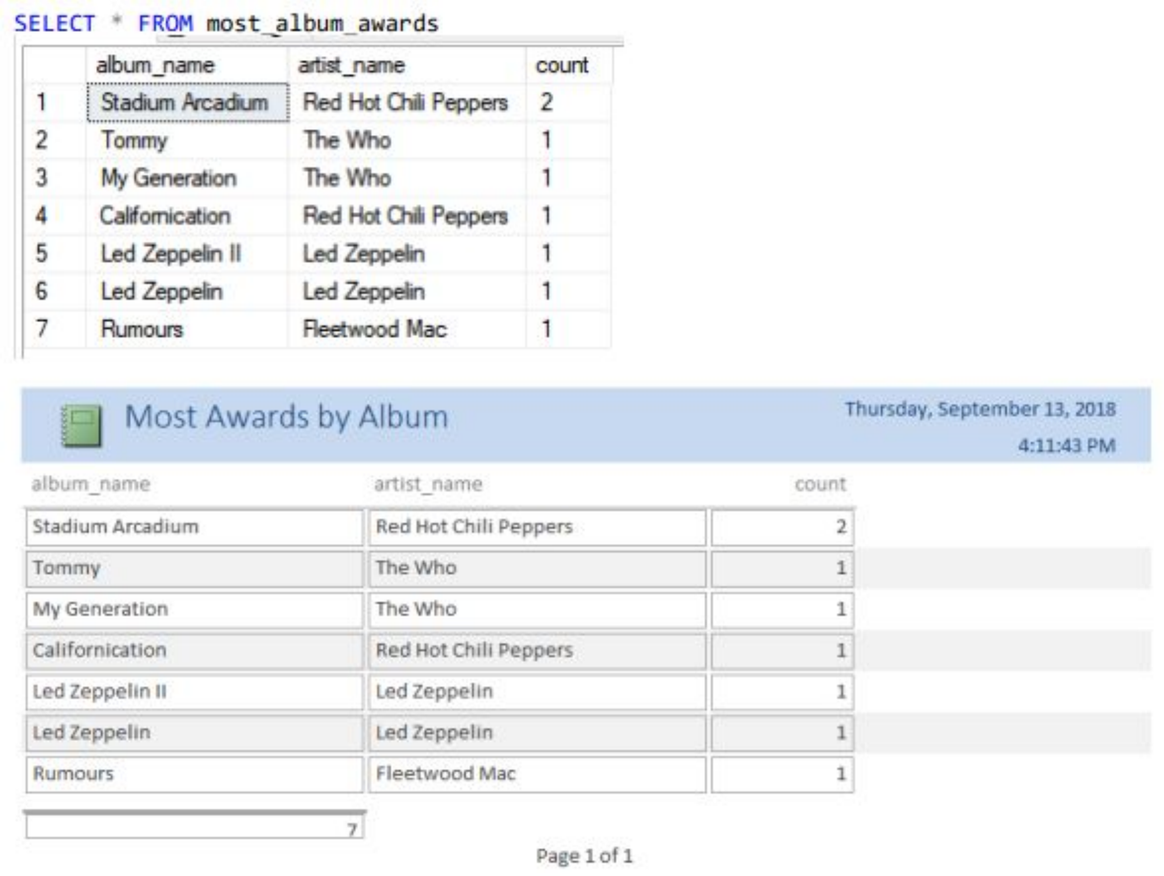
Most Awards by Artist

Thursday, September 13, 2018
4:11:05 PM

artist_name	count
Red Hot Chili Peppers	3
Ben Folds	2
The Who	1
Led Zeppelin	1
Fleetwood Mac	1

5

Page 1 of 1



```
SELECT * FROM most_album_awards
```

	album_name	artist_name	count
1	Stadium Arcadium	Red Hot Chili Peppers	2
2	Tommy	The Who	1
3	My Generation	The Who	1
4	Californication	Red Hot Chili Peppers	1
5	Led Zeppelin II	Led Zeppelin	1
6	Led Zeppelin	Led Zeppelin	1
7	Rumours	Fleetwood Mac	1

Most Awards by Album

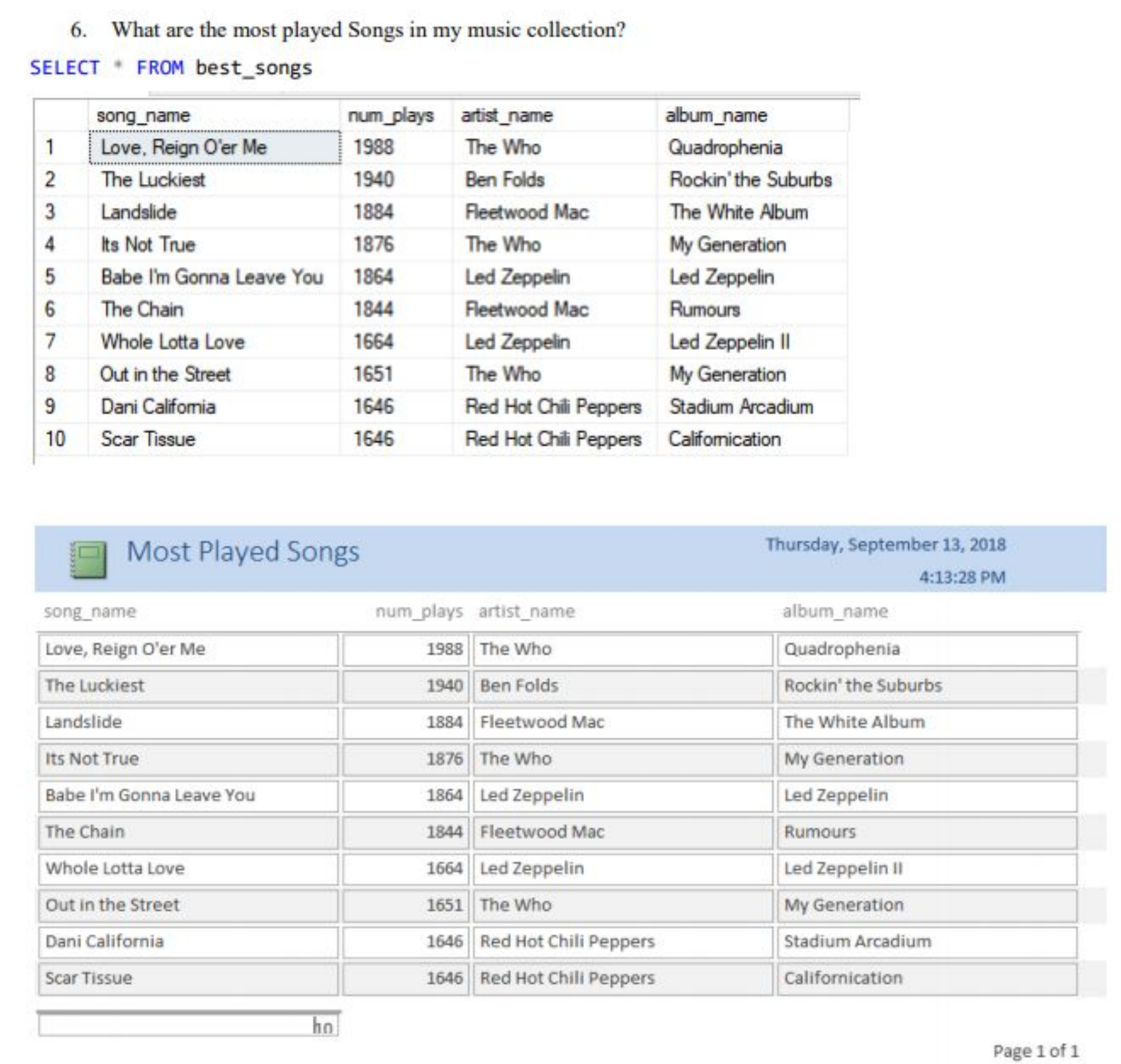
Thursday, September 13, 2018
4:11:43 PM

album_name	artist_name	count
Stadium Arcadium	Red Hot Chili Peppers	2
Tommy	The Who	1
My Generation	The Who	1
Californication	Red Hot Chili Peppers	1
Led Zeppelin II	Led Zeppelin	1
Led Zeppelin	Led Zeppelin	1
Rumours	Fleetwood Mac	1

7

Page 1 of 1

IST 659: Database Administration



6. What are the most played Songs in my music collection?

```
SELECT * FROM best_songs
```

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

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

Page 1 of 1

IST 659: Database Administration

Reflection

- 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.
- This application is applicable to the field of marketing analytics which often requires precise and efficient information management in order to deliver actionable insights.
- Operational marketing tasks such as promotional tagging are easily automated with the skills developed in this application using tools such as SSIS.
- In the final term of the program, additional study in more advanced database architectures will be explored in Advanced Database Management and Data Warehouse.

IST 707
Data Analytics:

Mushroom Classification

Poisonous or Edible? [5]

IST 707: Data Analytics

Introduction

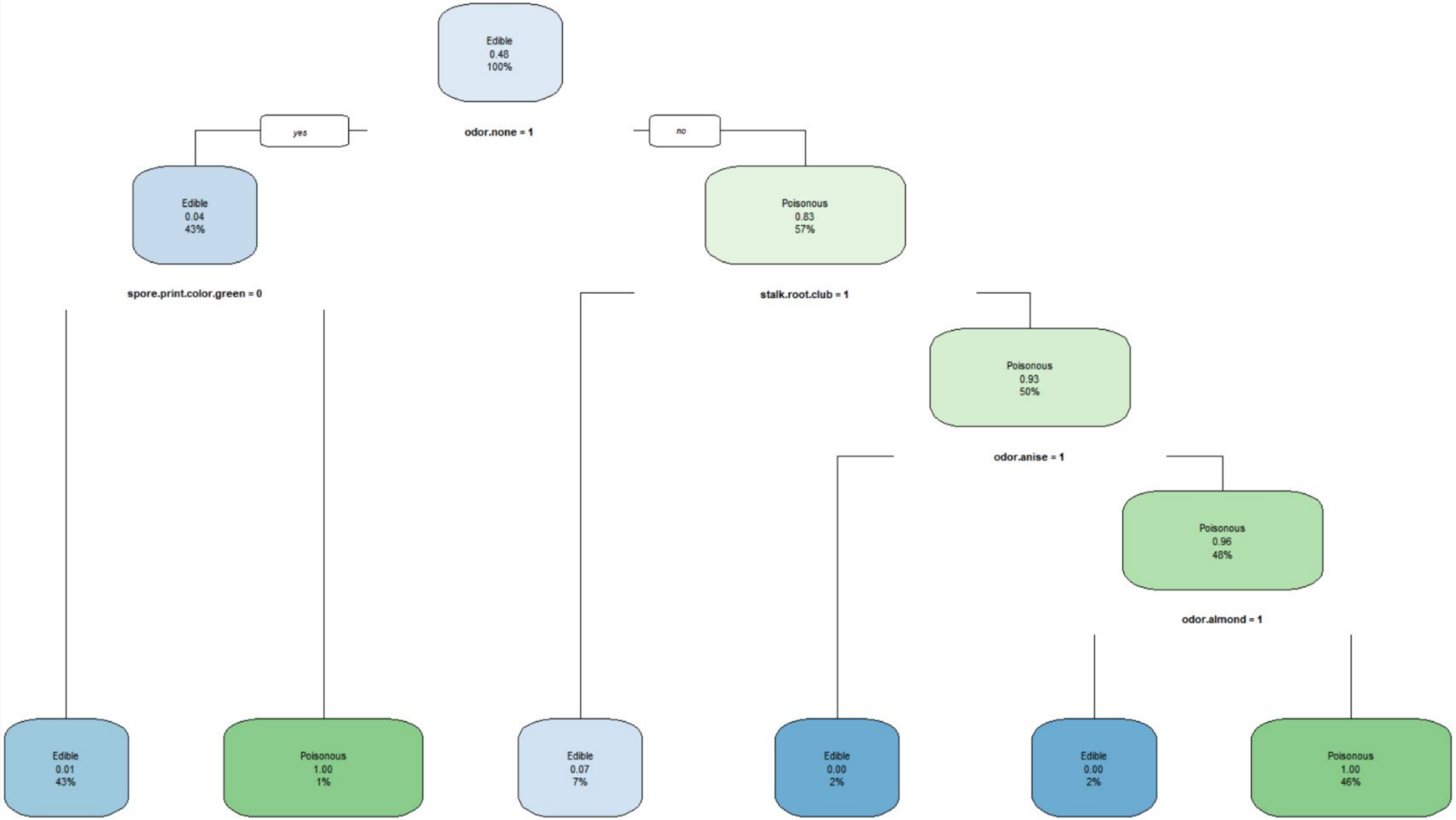
- Through studying Data Analytics under the direction of Dr. Mohammed Syed, various data mining techniques were introduced which perform with varying precision and efficiency for applications in regression, classification, and clustering.
- In this application, Naive Bayes and Decision Tree Classifiers are compared with respect to computation time and accuracy to predict the edibility of mushrooms.
- R Studio is leveraged to conduct analysis using data from Kaggle.

https://github.com/SLPeoples/MSADS_Portfolio/blob/master/IST707_DataAnalytics

Methodology

- This project required the cleaning and preprocessing of data.
    - This involved the transformation of 22 categorical features into 96 boolean vectors.
    - One-Hot Encoding also facilitated the identification of patterns within the data.
- Recent findings suggest Naive Bayes is the preferred classification algorithm over decision trees.[2,8]
- However, observations made in this application revealed nearly three percent greater accuracy across five-fold cross-validation using decision trees in one-tenth the time in comparison to Naive Bayes.
    - Decision trees predicted with an average 99.14% accuracy in 1.7 seconds of training.
    - Naive Bayes predicted with an average 96.51% accuracy in 13.9 seconds of training.

IST 707: Data Analytics



IST 707: Data Analytics

Reflection

- 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.
- This is an important distinction in a marketing analytics setting, which 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.

IST 736
Text Mining:

Congressional Twitter Analysis

LDA Topic Modeling &
K-Means Clustering of Member Tweets [6]

https://github.com/SLPeoples/MSADS_Portfolio/blob/master/IST736_TextMining

IST 736: Text Mining

Introduction

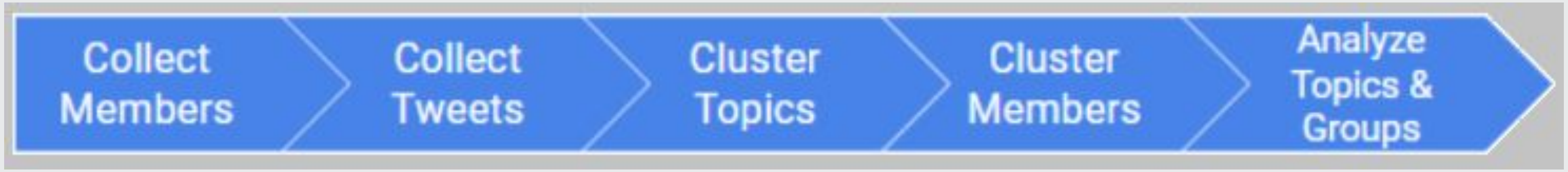
- Through studying Text Mining under the direction of Dr. Bei Yu, data mining techniques were introduced to analyze text and develop insights from unstructured data.
- In this application, Latent Dirichlet Allocation is leveraged in tandem with K-Means Clustering to group Congressmen by the content of their Tweets.
- This provides insight into the policies by which groups of lawmakers stand.
- Python is used for data collection and clustering, while Mallet is used to conduct LDA topic modeling. Exploratory data analysis is conducted using Tableau.

Methodology

- This technique was also successfully implemented using unstructured data in Bhoi’s research, which classified healthy and unhealthy gaits in Parkinson’s patients.[1]

- Miha Pavlinek also found success using LDA to improve classification tasks using Mallet.[3]

- The project was completed in five steps:



Collect Members

Collect Tweets

Cluster Topics

Cluster Members

Analyze Topics & Groups

IST 736: Text Mining

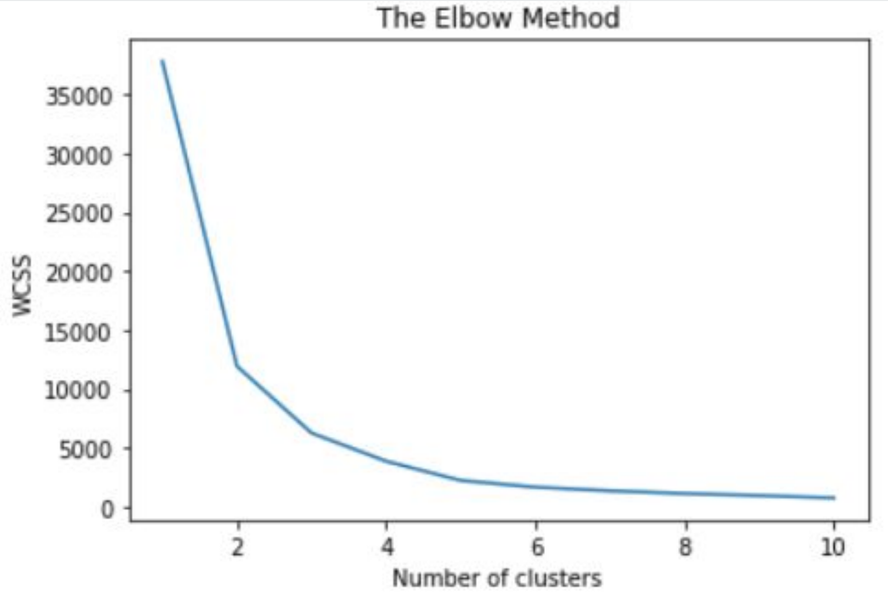
Methodology

- 527 of 535 members of Congress were associated with active Twitter accounts through a list managed by CSPAN.
- The latest 200 Tweets from each account were requested and saved to unique text files.
- This corpus of Tweets were then read into Mallet, testing & inspecting 10, 15, 25, and 50 clusters.
- 50 clusters were selected to account for the most variance between topics.
- The LDA weights were then paired with each member's party affiliation, house seat, and seniority in years before being clustered.

IST 736: Text Mining

Methodology

- The elbow method is used to observe an optimal number of clusters of $k = 3$.
- Following the secondary clustering, the data is exported to Tableau and visualized.

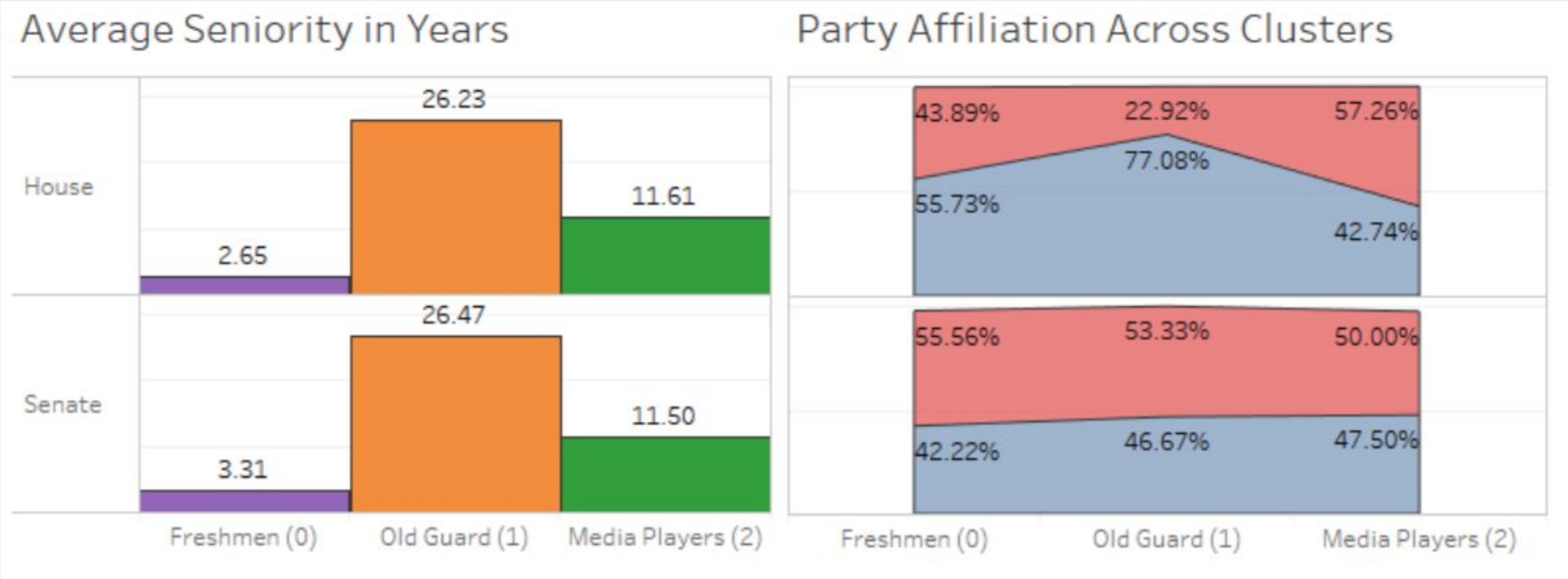


IST 736: Text Mining

Profiling

- Three distinct groups are identified:
  - Cluster 0: Freshmen - Congress members with very little experience. Slightly more Republican Senators, similar distribution for Democrat House members.
  - Cluster 1: Old Guard - Members with the most experience. Primarily Democrats in the House, slight majority towards Republicans in the Senate.
  - Cluster 2: Media Players - Lawmakers who have a ranging seniority and discuss more rhetoric-filled topics. Primarily Republicans in the House and Senate.
- Independents are excluded from profiling, and is represented in the visualization’s percentages.

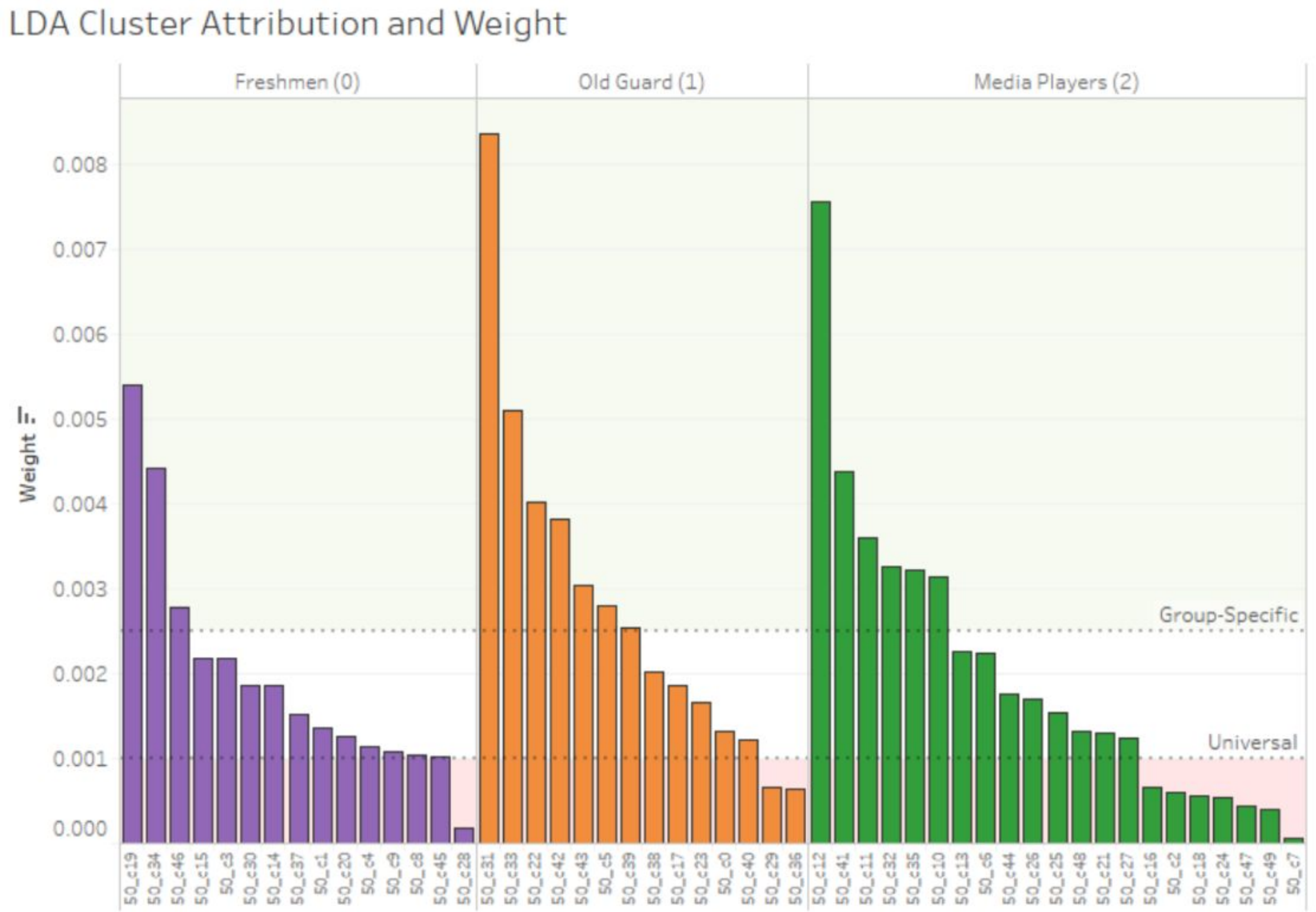
IST 736: Text Mining



Cluster-Topic Attribution

- Attribution of each topic cluster is applied to each K-Means cluster.
- Weights are calculated by averaging the LDA weights across each cluster and subtracting that from the average of the within-cluster weights.
- This provides the magnitude of weight that is attributed to each topic.
- Universal topics are identified as topics with weight less than 0.001.
- Biased group-specific topics have weights greater than 0.0025.
    - Ten universal topics and sixteen group-specific topics are identified.

IST 736: Text Mining



https://github.com/SLPeoples/MSADS_Portfolio/blob/master/IST736_TextMining

IST 736: Text Mining

Reflection

- 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.
- User privacy was considered to both request only the necessary data and maximize the API rate limitations.
- In a marketing analytics setting, the ability to manage and analyze text data is growing in significance as organizations project a larger presence in social media.
  - Text mining can be leveraged against large collections of text which allows for automation using conversational assistants, as well as predictive analytics..

MAR 653
Marketing Analytics:

Complete Journey

K-Means Clustering & Apriori Rule Association
For Development of Direct Mail Promotion [7]

MAR 653: Marketing Analytics

Introduction

- Through studying Marketing Analytics under the direction of Dr. Andrew Petersen, data mining concepts specific to marketing were introduced which inspired the final presentation.
- Yearly transaction information for households from a grocery store was used to identify a target group of customers using K-Means Clustering.
- An optimal promotional offer was then derived using Apriori Rule Association and Sensitivity Analysis for use in a Direct Mail Marketing Campaign.
- Python was used for clustering and rule association, while Sensitivity Analysis and data exploration were conducted with Excel.

https://github.com/SLPeoples/MSADS_Portfolio/blob/master/MAR653_Marketing_Analytics

MAR 653: Marketing Analytics

Segmentation

- This exercise involved exploring and cleaning the data prior to clustering.
  - Categorical columns are one-hot encoded, dropping one to avoid the dummy-variable trap.
- Segmentation is completed using the items and pricing data, while profiling is accomplished using demographic information such as:
  - Age
  - Income
  - Household Size
  - Marital Status
  - Number of Children
- Considerations are made to exclude demographic data from segmentation to avoid introducing bias into the analysis.

https://github.com/SLPeoples/MSADS_Portfolio/blob/master/MAR653_Marketing_Analytics

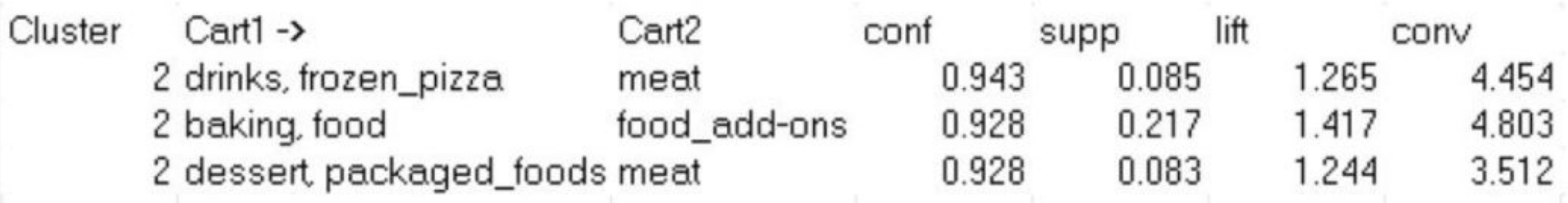
MAR 653: Marketing Analytics

Optimal Cart Building

- The selected cluster contained nearly 20% of the entire customer base and use coupons slightly more often than the average customer.
- Three optimal carts are developed using Apriori Rule Association and the items purchased by households in this cluster.
- These items are selected such that when items from *cart 1* are purchased, items from *cart 2* have a discount applied.
- This method has also been successfully implemented in predicting purchase behavior in an e-commerce setting, which also uses a target subset of their customers to improve the expected result.[9]

MAR 653: Marketing Analytics

Optimal Discount Calculation



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

- The optimal coupon discount is calculated using the average discounts applied previously and varying the percent discount to not exceed a $10,000 liability margin.
- This resulted in an expected 137% increase in gross revenue for the affected products, with a $4,000 reduction in liability from other recurring promotions.
- The promotion was recommended to be ran for one month to measure the participation from the targeted group, with well defined goals and expectations to measure success.

MAR 653: Marketing Analytics

Reflection

- 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 necessary to develop a plan of action to quantify the insights developed in this analysis, which translates to actionable business decisions.
- This application allowed data to guide the analysis, requiring alternative strategies to be developed as observations were made within the data.
- Ethical considerations were made with respect to segmentation, using demographic information to profile previous behavior.

The Applied Data Science Program has seven Learning Objectives:

1. Describe a broad overview of the major practice areas in 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 relevant professionals in their organization.
7. Synthesize the ethical dimensions of data science practice.

Learning Objectives

- This portfolio has demonstrated successful implementation of the program’s learning objectives and the major practice areas of data science. Each application involved an increasing level of complexity, and characterized the progression of these skills.
- Data was collected and managed using web scraping and APIs in conjunction with database solutions to be analyzed using statistical methods and data mining techniques for tasks such as regression, classification, or clustering.
  - IST 659 collected and managed data using database solutions.[4]
  - IST 707 analyzed data using Naive Bayes and Decision Tree classification techniques.[5]
  - IST 736 clustered phrases and sources using Latent Dirichlet Allocation and K-Means Clustering.[6]
  - MAR 653 utilized K-Means Clustering, Apriori Rule Association, and Sensitivity Analysis.[7]

Learning Objectives

- Various Data Visualizations were paired with clustering techniques to identify patterns which directed the respective analyses; actionable recommendations were developed to reflect tangible business decisions. As a result, it was necessary to develop alternative methods based on the outcome of each step in the analyses.
    - IST 736 was directed by the results of the LDA clustering which revealed patterns in the content of Congressional members' Tweets. The analysis was then leveraged to deliver actionable insights regarding how topics discussed by lawmakers are categorized.[6]
    - MAR 653 was directed solely by the results of the analyses, where a target customer group was first identified, followed by the selection of frequently purchased items within this group, and finally the calculation of the suggested discounts based on these selected items. This allows for confident recommendations to be delivered which translate to tangible business decisions.[7]

Learning Objectives

- Communications skills were displayed in the organization and delivery of insights, expressing them in terms which could be simply understood and acted upon.[4,5,6,7]
- The ethical dimensions of data science practice were displayed and considered when analyzing PII from external sources.
  - IST 736 required the consideration of user privacy when requesting Tweets from the Twitter API. Only the relevant information was requested to balance both rate limitations and user privacy.[6]
  - MAR 653 required the consideration of demographic information which was excluded from segmentation, being used only for profiling. This reduces the possibility of introducing bias into the analysis.[7]
  - Additional study of the ethical dimensions of data science will take part while taking Information Policy in the final term of the program. Topics such as intellectual property and freedom expression will be explored.
- These projects are representative of the successful execution of the learning objectives and have developed the necessary skills for practice in the field of data science.

Conclusion

- Syracuse University’s School of Information Studies provides students the opportunity to synthesize the collection, management, and analysis of data, as well as the 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 with increasing complexity.
- It has cultivated strategies which improve organizational efficiency.
- The program has fostered a practice of transparency, reproducibility, and ethical data management which promotes integrity within an organization’s analytics team.
- Using the skills learned in the Applied Data Science program, 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.

Thank You!

Template Retrieved from:
https://www.syracuse.edu/about/brand/visual-identity/powerpoint-templates/

References

1. 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

2. 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

3. 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

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

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

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

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

8. 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

9. 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

Image Sources

A. *Kaggle*[PNG]. (n.d.). https://aezmdcb0d81bgava-zippykid.netdna-ssl.com/wp-content/uploads/2017/03/Kaggle-logo.png.

B. *Mallet*[PNG]. (n.d.). http://mallet.cs.umass.edu/logo3.png.

C. *Microsoft Access*[JPG]. (n.d.). https://www.microsoft.com/en-us/microsoft-365/blog/wp-content/uploads/sites/2/migrated-images/53/8787.CAROUSEL_Access_260x146.jpg.

D. *Microsoft Excel*[PNG]. (n.d.). https://gravitatesolutions.com/refresh/wp-content/uploads/2015/09/connector-excel-logo.png.

E. *Python*[PNG]. (n.d.). https://www.python.org/static/community_logos/python-logo-master-v3-TM.png.

F. *R Studio*[PNG]. (n.d.). https://www.dcu.ie/sites/default/files/software/r_studio.png.

G. *SciKit Learn*[PNG]. (n.d.). https://scikit-learn.org/stable/_images/scikit-learn-logo-notext.png.

H. *SQL Server*[SVG]. (n.d.). https://cdn.worldvectorlogo.com/logos/microsoft-sql-server.svg.

I. *Tableau*[PNG]. (n.d.). https://www.sheerid.com/wp-content/uploads/2016/10/tableau-logo.png.

J. *Twitter*[JPG]. (n.d.). Https://static01.nyt.com/images/2014/08/10/magazine/10wmt/10wmt-articleLarge-v4.jpg?quality=75&auto=webp&disable=upscale.