PORTFOLIO MILESTONE

ASHRAF WAN

SUID: 273192117

TABLE OF CONTENT

- Introduction
- IST 707 – APPLIED MACHINE LEARNING
- IST 719 – INFORMATION VISUALIZATION
- IST 736 – TEXT MINING
- MBC 638 – DATA ANALYSIS & DECISION MAKING
- CONCLUSION

INTRODUCTION

- The Applied Data Science program at Syracuse University's School of Information Studies is designed to have students master the fundamental aspects of data science through project-based research and deliverables. Each student is to demonstrate mastery in data collection, data analysis, and implement business decisions. Through this portfolio, I will showcase my understanding of the different aspects of data science through different courses and different tools.

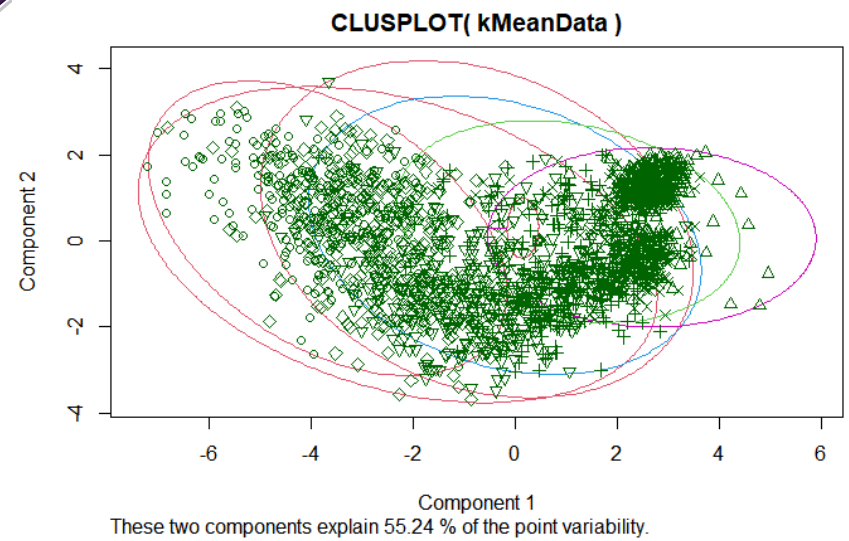
IST 707 – APPLIED MACHINE LEARNING

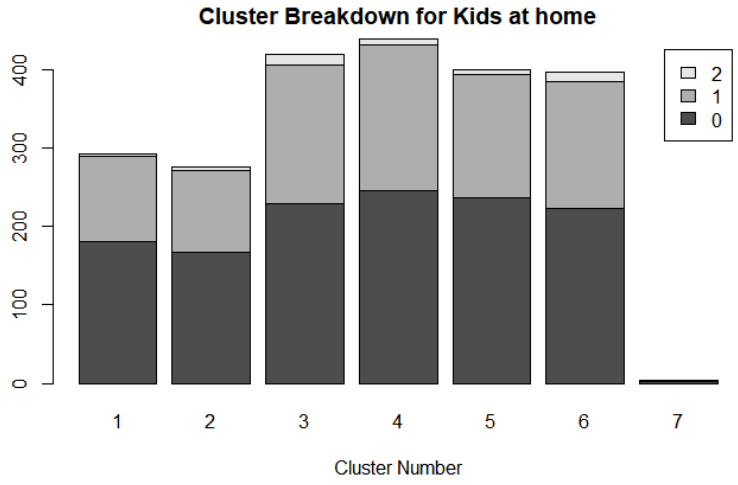
- Purpose:
  - perform a customer personality analysis based on a grocery store's dataset and find the ideal customer to target for a marketing campaign.
- Dataset:
  - Grocery Store's Customer data (Kaggle)
- Technology:
  - R
- Techniques:
  - Apriori Algorithm
  - K-Means Clustering
  - Decision Tree
  - Naïve Bayes
  - Random Forest
  - SVM
  - KNN

APRIORI ALGORITHMS

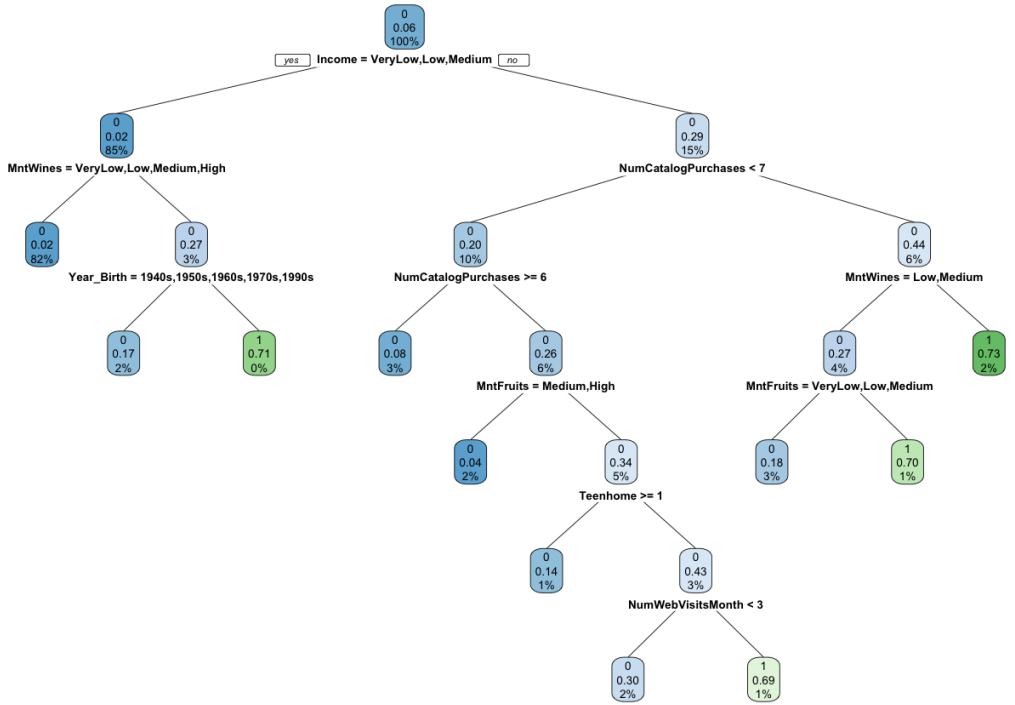
```
Apriori
=======

Minimum support: 0.2 (448 instances)
Minimum metric <confidence>: 0.1
Number of cycles performed: 16

Generated sets of large itemsets:

Size of set of large itemsets L(1): 9

Size of set of large itemsets L(2): 6

Best rules found:

 1. Kidhome='(0.666667-1.333333]' Teenhome='(-inf-0.666667]' 503 ==> AcceptedCmp1='(-inf-0.5]' 498    conf:(0.99)
 2. Kidhome='(0.666667-1.333333]' 899 ==> AcceptedCmp1='(-inf-0.5]' 890    conf:(0.99)
 3. Teenhome='(0.666667-1.333333]' 1030 ==> AcceptedCmp1='(-inf-0.5]' 1003    conf:(0.97)
 4. Education=Graduation Teenhome='(0.666667-1.333333]' 511 ==> AcceptedCmp1='(-inf-0.5]' 497    conf:(0.97)
 5. Kidhome='(-inf-0.666667]' Teenhome='(0.666667-1.333333]' 625 ==> AcceptedCmp1='(-inf-0.5]' 603    conf:(0.96)
 6. Marital_Status=Together 580 ==> AcceptedCmp1='(-inf-0.5]' 548    conf:(0.94)
 7. Education=PhD 486 ==> AcceptedCmp1='(-inf-0.5]' 456    conf:(0.94)
 8. Marital_Status=Single 480 ==> AcceptedCmp1='(-inf-0.5]' 449    conf:(0.94)
 9. Education=Graduation 1127 ==> AcceptedCmp1='(-inf-0.5]' 1045    conf:(0.93)
10. Marital_Status=Married 864 ==> AcceptedCmp1='(-inf-0.5]' 801    conf:(0.93)
11. Teenhome='(-inf-0.666667]' 1158 ==> AcceptedCmp1='(-inf-0.5]' 1043    conf:(0.9)
12. Kidhome='(-inf-0.666667]' 1293 ==> AcceptedCmp1='(-inf-0.5]' 1160    conf:(0.9)
13. Education=Graduation Teenhome='(-inf-0.666667]' 593 ==> AcceptedCmp1='(-inf-0.5]' 527    conf:(0.89)
14. Education=Graduation Kidhome='(-inf-0.666667]' 650 ==> AcceptedCmp1='(-inf-0.5]' 575    conf:(0.88)
15. Kidhome='(-inf-0.666667]' Teenhome='(-inf-0.666667]' 638 ==> AcceptedCmp1='(-inf-0.5]' 528    conf:(0.83)
```

K-MEANS CLUSTERING



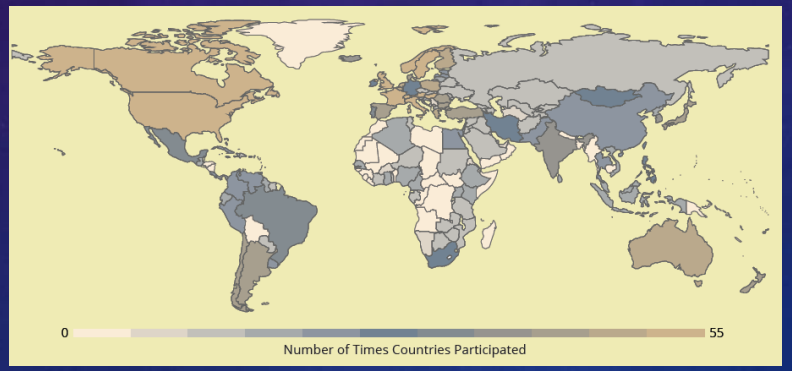


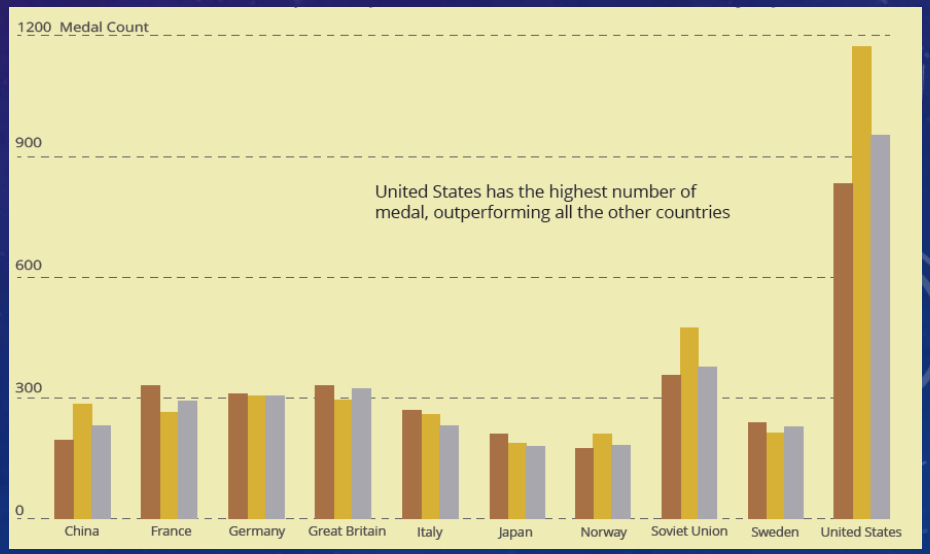
DECISION TREE



IST 719 – INFORMATION VISUALIZATION

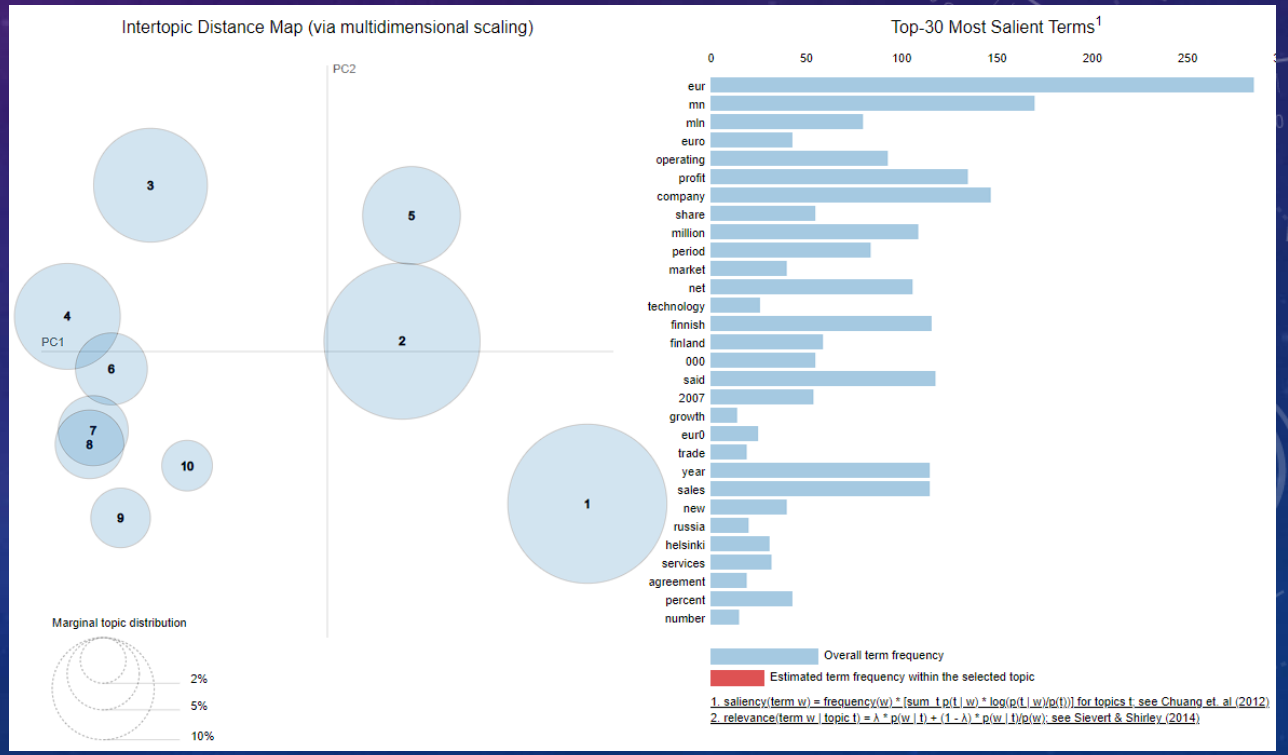
- Purpose:
  - Take a dataset and create a poster to showcase the results of the analytics through visualization.
- Dataset:
  - Every country's Olympic game's performance for both Summer and Winter Olympic (Kaggle)
- Technology:
  - R
  - Adobe Illustrator
- Techniques:
  - ggplot
  - rworldmap





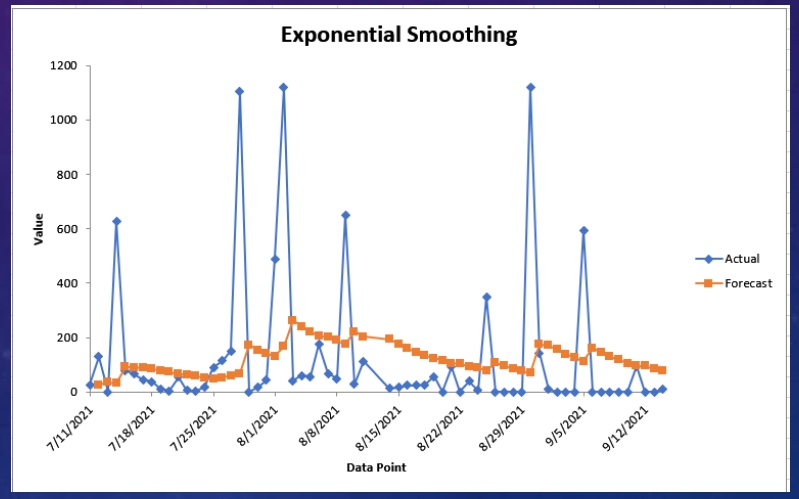
IST 736 – TEXT MINING

- Purpose:
  - Take a dataset and perform sentiment analysis and topic modeling.
- Dataset:
  - Financial news headline from 2014 (Kaggle)
- Technology:
  - Python
- Techniques:
  - Naïve Bayes
    - Multinomial
    - Bernoulli
  - SVM
  - Topic Modeling



MBC 638 – DATA ANALYSIS & DECISION MAKING

- Purpose:
  - Showcase understanding of various analytics technique and excel usage.
- Dataset:
  - Improvement of my monthly savings (My personal bank statements)
- Technology:
  - Excel
- Techniques:
  - Descriptive Statistics
  - Chi-squared Test
  - Correlation
  - Exponential Smoothing



	Bills and Debts	*Groceries*	*Take Out Food*	*Essentials*	*Random*	*Total*
Bills and Debts	1					
Groceries	-0.015405492	1				
Take Out Food	0.050644291	0.064033714	1			
Essentials	-0.084653834	-0.076510929	0.005868123	1		
Random	0.136485673	-0.07420402	-0.108178086	-0.062099937	1	
Total	0.879917329	-0.00417998	0.052340228	-0.04541087	0.581658801	1

Spending	
Mean	97.88982143
Standard Error	27.7212269
Median	26.265
Mode	94.99
Standard Deviation	207.4466668
Sample Variance	43034.11957
Kurtosis	12.7393003
Skewness	3.476666464
Range	1117.14
Minimum	2.86
Maximum	1120
Sum	5481.83
Count	56

CONCLUSION

- This portfolio has demonstrated the various implementation of the learning objectives using different tools from Python to R to Excel and to and outdated tool such as WEKA. Majority of the data were collected through Kaggle, and any missing information is filled in using other sources. The datasets were analyzed using statistical methods and data mining techniques such as clustering and classification and presented using visualizations. Not only were these skills showcased for individual projects, but these were also showcased successfully in a group environment as well.

https://github.com/ashwan01/datascience-portfolio