Applied Data Science: Final Portfolio

Emma Gillen

IST 623

Information Security

Overview

- Atlanta ransomware attack in March 2018
- The attack on Atlanta is considered the largest, most expensive cyber disruption in city government to date
- Orchestrated by the SamSam group
  - Encrypted files
  - Displayed a ransom note
  - Demanded a Bitcoin ransom in exchange for a key to decrypt the affected computers

The Attack

- Gained access to these computers using vulnerabilities in their web applications and servers
  - The cyber actors had purchased several stolen RDP credentials from darknet marketplaces
- Escalate their own privileges to those of an administrator and dropped malware on the server
- The ransom could be paid domain that was only accessible using a TOR browser
- After paying, they would receive a link to download the decryption keys
  - Paying a ransom does not guarantee results
  - The decision whether or not to pay a ransom requires evaluation of all alternatives

Vulnerabilities & Consequences

- The attack highlighted more than 2,000 major vulnerabilities in the Atlanta municipal system
- An audit performed three months prior to the attack showed there were known issues
  - A number of recommendations were made in the audit that had not been put in place
- The attack caused millions of dollars in damages
  - Estimates ranged from $7 million to $17 million
  - Lost productivity as city employees were unable to use their computers for five days

Objectives Covered

Objective 7 - Synthesize the ethical dimensions of data science practice (e.g., privacy).

- Data needs to be protected for a variety of reasons depending on the specific situation
  - Consumer: concerns about sensitive, personal data being spread
  - Company: concerns about industry information being given to competitors
- Possible Solutions:
  - Encryption
  - Permissions
- Atlanta ransomware attack highlighted the importance of data security
  - Millions of dollars in damages and halted many essential operations
  - Professionals were aware of these issues and could have potentially prevented the attack

IST 719

Information Visualization

goodreads

judging a book by its ~~cover~~ DATA

Emma Gillen IST 719

Story: The goodreads website contains entries on thousands of books as well as user reviews and ratings. Based on this user feedback, goodreads provides recommendations.

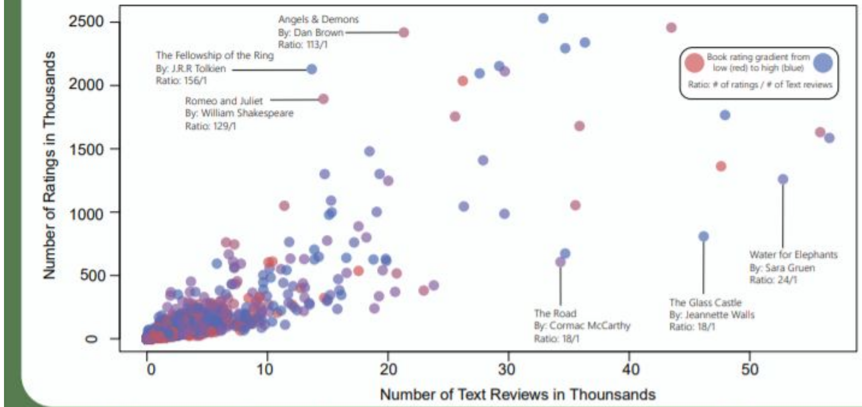
Motivation: Popular or trending media is often reccomended. However, there is an often a conflict between popularity and quality. This study focuses on what combination of factors can be considred to find a "good read".

The Data: The dataset contains 11,131 rows and 12 columns. For analysis, books with less than 10 reviews (too obscure) or less than 100 pages in length (largely childrens book and incorrect entries) were removed, leaving 9,584 rows. The key columns used for analysis include the title, author, average rating, number of ratings, number of written reviews, and page length.

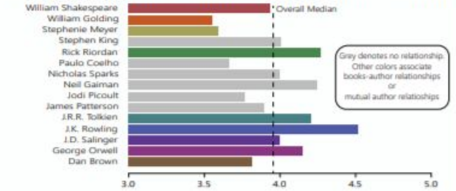
Does the average rating have a relationship with the amount of feedback a book recieves? An author?

Review & Rating Counts by Score

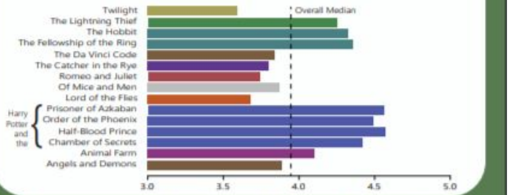
All zero values have been removed in addition to two outliers. These two books are Twilight (Ratio: 49/1) and The Book Thief (Ratio: 17/1).



Average Rating for the Top Fifteen Most Rated Authors



Average Rating for the Top Fifteen Most Rated Titles

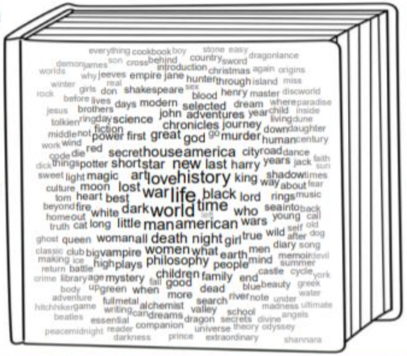


What are these books about?

Most Common Words in Book Titles

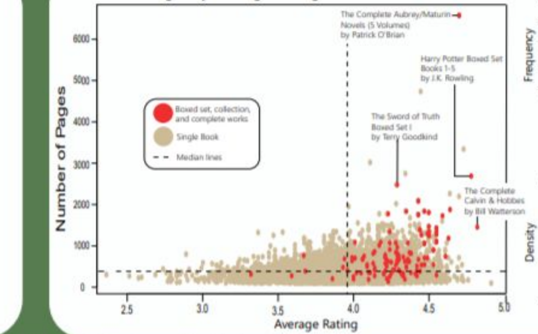
Function words, or words with little meaning were filtered from the word cloud.

Words that occured less than twenty times were filtered.

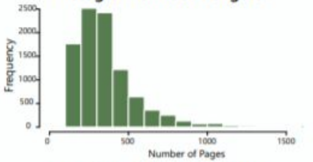


Does the page length have a relationship with rating?

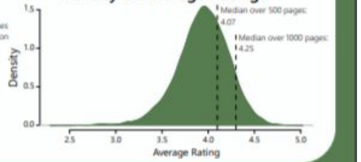
Book Length by Average Rating



Histogram of Book Lengths



Density of Average Ratings

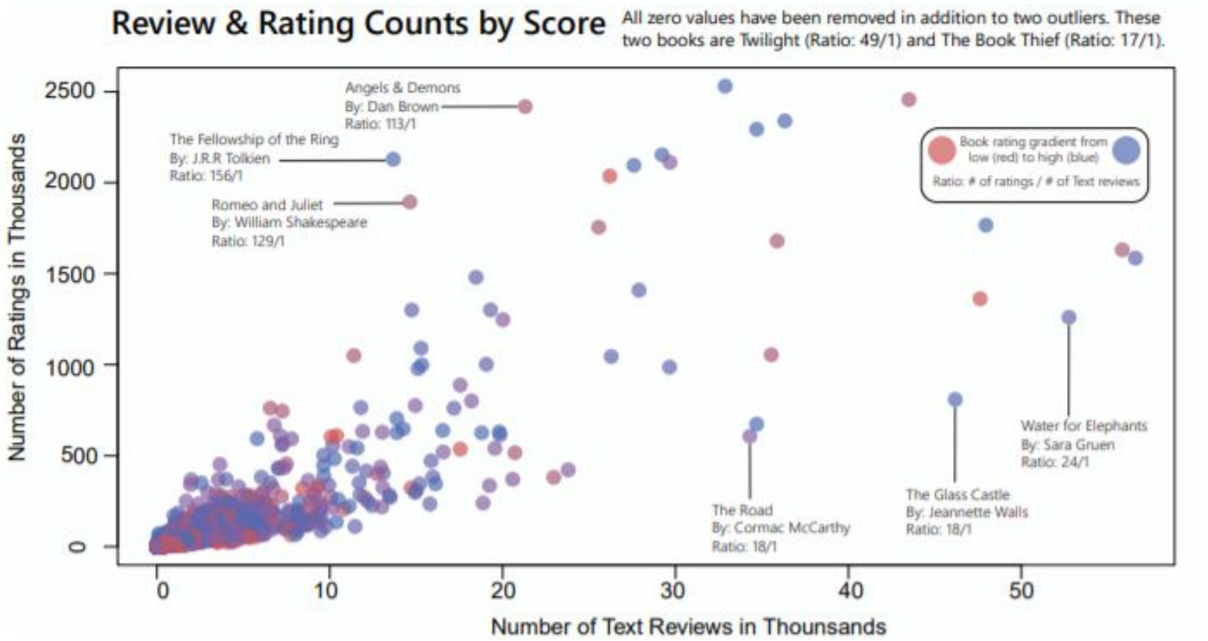


Sources: https://www.kaggle.com/jealousleopard/goodreadsbooks, https://www.kindpng.com/picc/m/84-843028_book-clipart-square-blank-book-cover-clip-art.png

R Packages: stringr, plyr, scales, wordcloud, tidytext

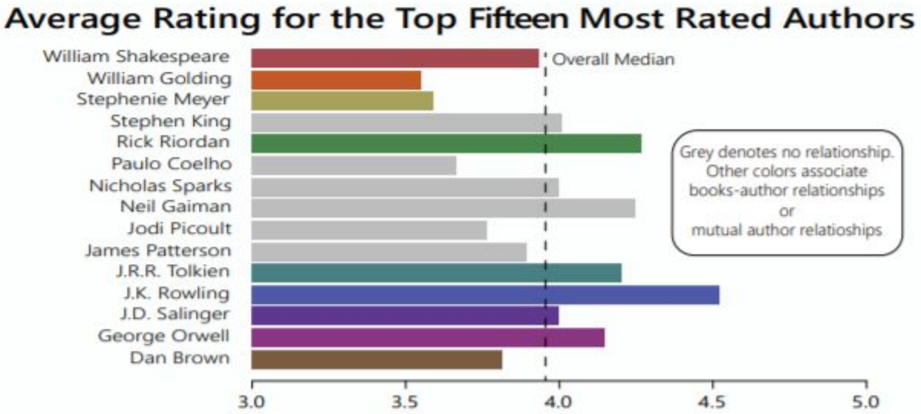
Scatterplot - Does the average rating have a relationship with the amount of feedback a book receives? An Author?

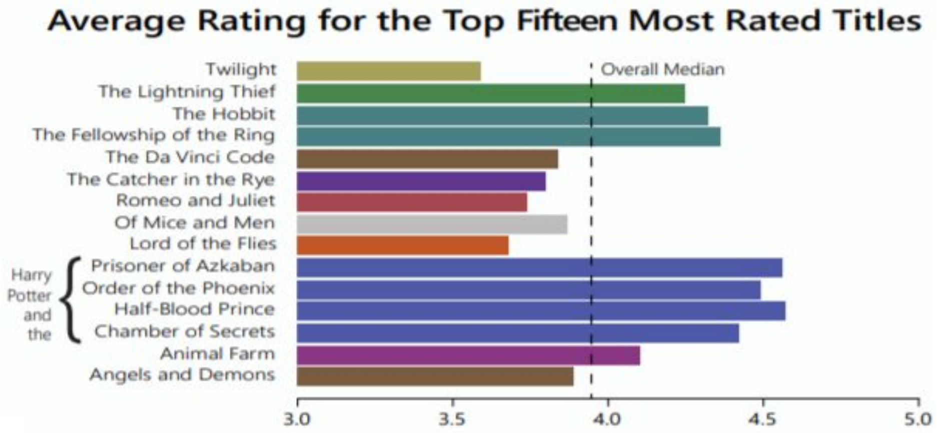
- Positioned in the upper left corner where the reader's eye would first be drawn
- Visualizes the relationship between the two forms of user feedback
- Six outliers called out with their specific book information
- A color gradient from red to blue indicates the average book rating



Barplots - Does the average rating have a relationship with the amount of feedback a book receives? An Author?

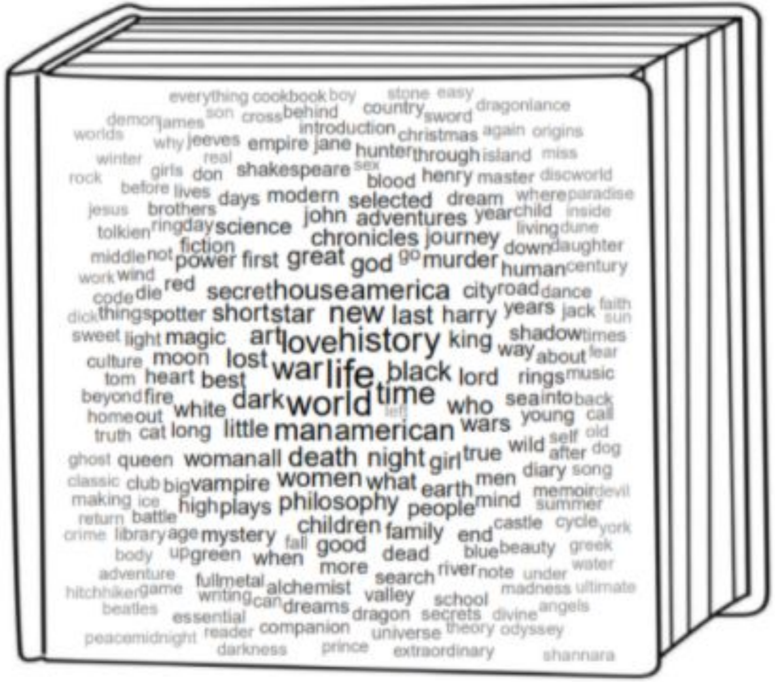
- Displayed in the upper righthand corner of the poster
- Data aggregated to highlight key information
- A dashed line is used to indicate the median rating in both plots
- The colors of the bars are used to associate books with the same author or book-to-author relationships





Word Cloud - What are these books about?

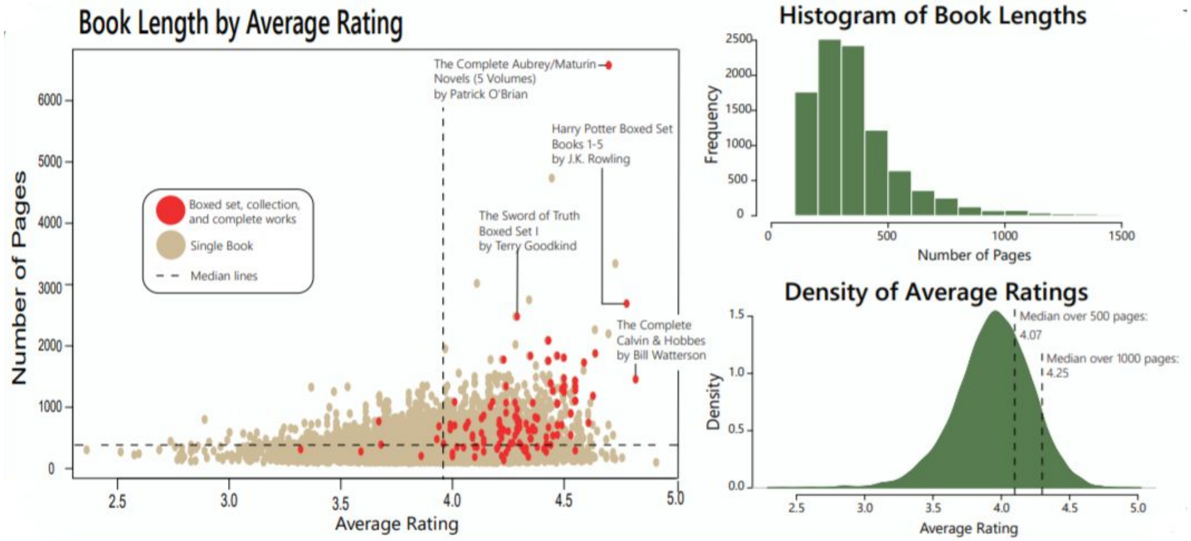
- A word cloud was generated using the most commonly used words in the book titles
- Displayed in the lower left hand corner of the original poster
- The word cloud was used to give a high level overview of the selection of books in the Goodreads dataset.
- The words displayed are filtered of any function words that do not add context to the titles



everything cookbook boy stone easy
demon james son cross behind country sword dragonlance
worlds introduction christmas again origins
why jeeves empire jane hunter through island miss
winter real
girls don shakespeare sex
rock blood henry master discworld
before lives days modern selected dream where paradise
jesus brothers john adventures year child inside
tolkien ring day science living dune
fiction chronicles journey
middle not power first great god go murder down daughter
work wind human century
red
code die secret house america city road dance
dick things potter short star new last harry years jack faith sun
sweet light magic art love history king shadow times
culture moon lost way about fear
tom heart best war life black lord rings music
beyond fire dark world time who sea into back
home out white left young call
truth cat long little man american wars old
self
ghost queen woman all death night girl true wild after dog
classic club big vampire women what men diary song
making ice high plays philosophy earth mind memoir devil
return battle people summer
crime library age mystery children family end castle cycle york
fall good
body up green when dead blue beauty greek
adventure more search river note under water
hitchhiker game fullmetal alchemist madness ultimate
writing can dreams valley school
beatles essential dragon secrets divine angels
peace midnight reader companion universe theory odyssey
darkness prince extraordinary shannara

Scatterplot with Supporting Plots - Does the page length have a relationship with rating?

- A scatterplot displaying the number of pages on the y-axis and the average rating on the x-axis
  - Dashed lines show medians
  - Red points show boxed sets, collections, or completed works
- A histogram of book lengths
- A density plot of the average rating
  - Dashed lines illustrate the median ratings for books over 500 and 1000 pages in length



Objectives Covered

Objective 3: Identify patterns in data via visualization, statistical analysis, and data mining.

- Box plots, histograms, and density plots can be used to visualize skew, kurtosis, distribution, and outliers
  - The density plot highlighted the kurtosis
  - The histograms and density plot showed the distribution
- Scatterplots may be utilized to identify trends, significant outliers, and clustering
  - Scatterplots also utilized colors and median lines

Objectives Covered

Objective 6: Demonstrate communication skills regarding data and its analysis for managers, IT professionals, programmers, statisticians, and other relevant professionals in their organization.

- One approach to conveying information is through visualizations
- In IST 719 - the goal was to show a large amount of information without requiring a large amount of technical knowledge
  - Clear and legible fonts
  - Effective layouts
  - Color

IST 736

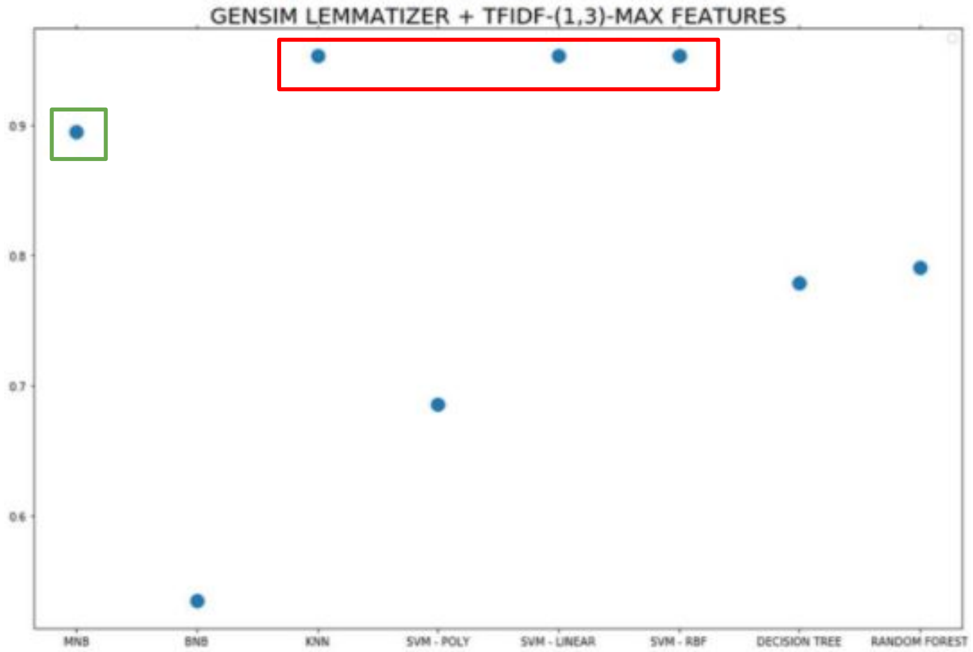
Text Mining

Overview

- An exploration of news articles to create a process to consume media devoid of bias
- Gathered articles from various sources with varying affiliations and perspectives and labeled:
  - Medical
  - Politics
  - Business
  - Sports
  - Entertainment
- Reduced dimensionality with either the Gensim lemmatizer or the Porter stemmer
- Vectorization with either binary and TFIDF vectorization
- Trained models on the labeled corpora

Model Accuracies

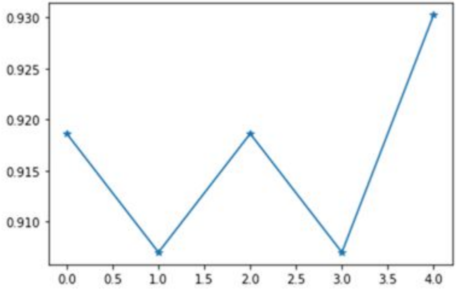
- 32 combinations of models and corpora
  - Multinomial and Bernoulli Naive Bayes
  - k-Nearest Neighbors (kNN)
  - Support Vector Machines (SVM) with various kernel
  - Decision trees and Random forests
- Gensim lemmatizer > Porter stemmer
- TFIDF vectorizer > binary vectorizer
- The best accuracies were found using SVM (linear kernel or radial kernel) or kNN
- Multinomial Naive Bayes is better suited
  - Time to run SVM and kNN is prohibitive
  - Feature ranking is is better suited

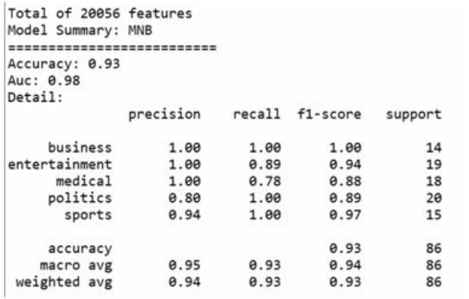


Multinomial Naive Bayes - Feature Selection

- Feature selection was explored
- 1st: MNB models were trained on different subsections of the data
  - Top 1000, 2000, 3000, 4000, and 5000 features for each label
  - The highest accuracy was 93%
- 2nd: chi square feature selection
  - The highest accuracy was 99%
  - Likely overfitted to our data
  - However, the feature selection is very indicative of each topic

Figure 1.3.9: Multinomial Naive Bayes: Top Feature Selection



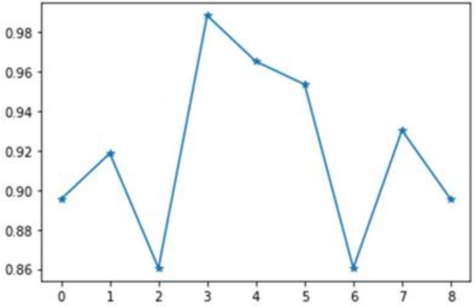


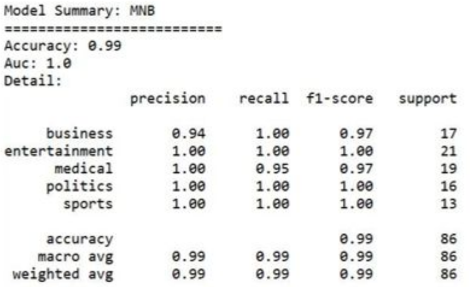
```
Total of 20056 features
Model Summary: MNB
==========================
Accuracy: 0.93
Auc: 0.98
Detail:
               precision    recall  f1-score   support

     business       1.00      1.00      1.00        14
entertainment       1.00      0.89      0.94        19
      medical       1.00      0.78      0.88        18
     politics       0.80      1.00      0.89        20
       sports       0.94      1.00      0.97        15

     accuracy                           0.93        86
    macro avg       0.95      0.93      0.94        86
 weighted avg       0.94      0.93      0.93        86
```

Figure 1.3.10: Multinomial Naive Bayes: P-value Manipulation





```
Model Summary: MNB
===========================
Accuracy: 0.99
Auc: 1.0
Detail:
               precision    recall  f1-score   support

     business       0.94      1.00      0.97        17
entertainment       1.00      1.00      1.00        21
      medical       1.00      0.95      0.97        19
     politics       1.00      1.00      1.00        16
       sports       1.00      1.00      1.00        13

     accuracy                           0.99        86
    macro avg       0.99      0.99      0.99        86
 weighted avg       0.99      0.99      0.99        86
```

Medical	Politics	Entertainment	Business	Sports
vaccine	trump	entertainment	stock	coach
health	election	actor	company	season
care	president	movie	investment	fantasy
study	republican	music	dividend	team
testing	senate	perry	oracle	football
disease	fraud	check fashion	business	sport
medicine	campaign	check fashion entertainment	analyst	fantasy football
test	court	division re	homology	game
drug	presidential	division re serve	price target	player
blood	trump campaign	entertainment gossip	income	fantasy football expert

Objectives Covered

Objective 1: Describe a broad overview of the major practice areas in data science.

- Data modeling: create a statistical model which can replicate or predict the behavior of the data
- Machine learning: trains a computer to improve the model automatically
- Models can be either supervised or unsupervised
  - Examples of supervised models: regression, support vector machines, and decision trees
  - Examples of unsupervised models: k-means clustering and association rule mining
- In IST 736 - the best choice of supervised model was selected by considering:
  - Accuracy
  - Time complexity
  - Type of output produced

Objectives Covered

Objective 2: Collect and organize data.

- Data cleaning and organization steps include:
  - Handling missing data
  - Removing any unnecessary data
  - Correcting data structuring
  - Correcting data types
- In IST 736 - an API was utilized to gather text data
  - Needs to be transformed into a computer understandable format

Objectives Covered

Objective 6: Demonstrate communication skills regarding data and its analysis for managers, IT professionals, programmers, statisticians, and other relevant professionals in their organization.

- Groups will include people with varying degrees of technical knowledge
- In IST 736 - an emphasis was placed on writing non-technical introductions and conclusions
  - Allows individuals to understand the high level problem and solutions without being overwhelmed with the technical aspects

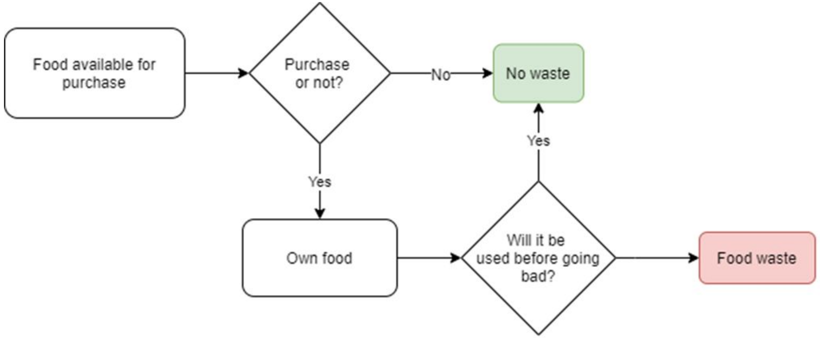
MBC 638

Data Analysis and Problem Solving

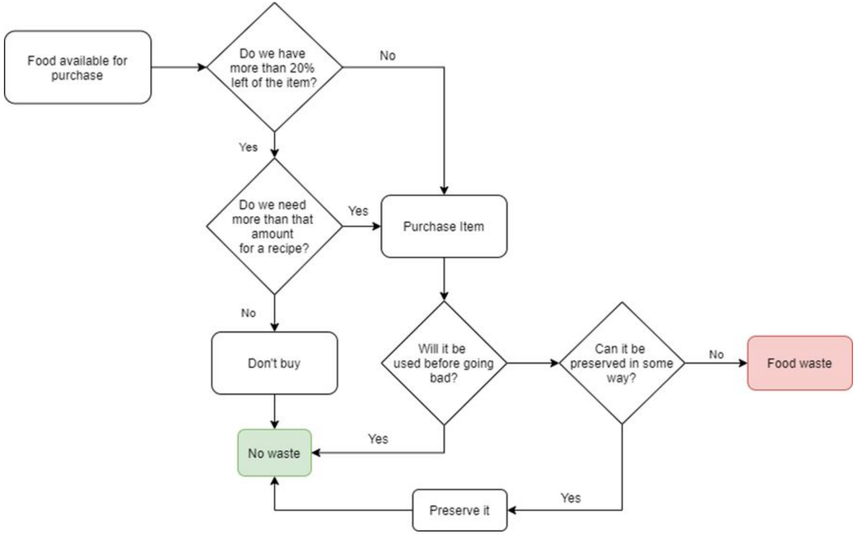
Process Improvement

- Food waste within my own household
- Original and improved processes are shown with flow charts
- Identified and implemented two changes

Before:

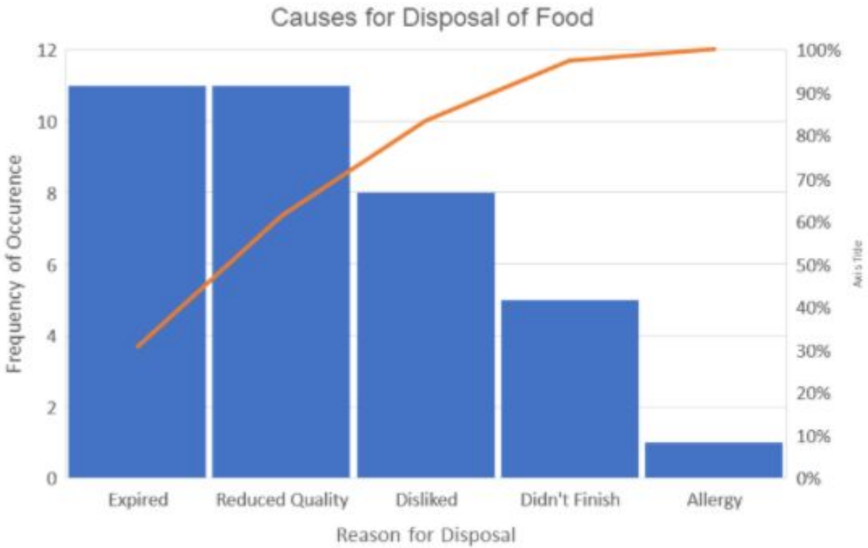

Food available for purchase
Purchase or not?
No
No waste
Yes
Yes
Own food
Will it be used before going bad?
Food waste

After:


Food available for purchase
Do we have more than 20% left of the item?
No
Yes
Do we need more than that amount for a recipe?
Yes
Purchase Item
No
Don't buy
Will it be used before going bad?
Can it be preserved in some way?
No
Food waste
Yes
No waste
Yes
Preserve it

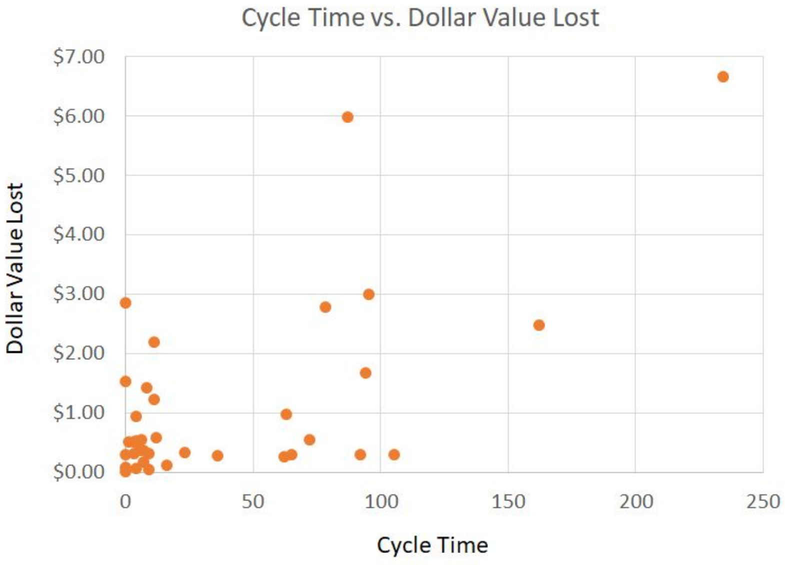
Food Waste Pareto Diagram

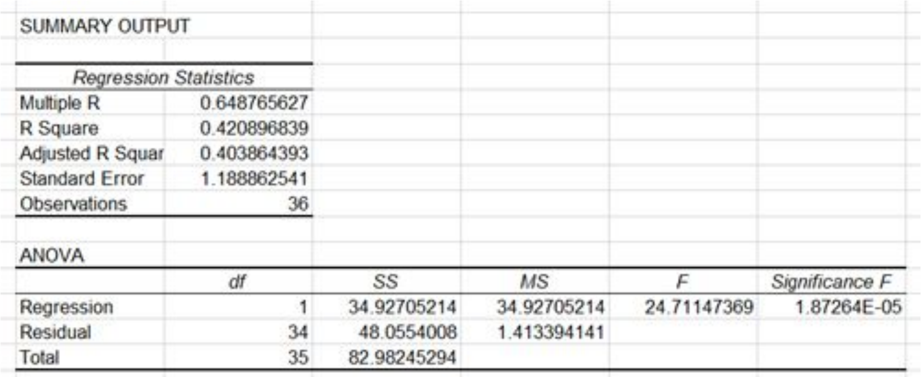
- The reason for disposal was recorded for each observation
  - One of five predetermined scenarios
- A pareto chart displaying the occurrence of these scenarios was created
  - 80% of food disposal took place in two of the five scenarios



Regression

- The relationship between the cycle time and the dollar value lost was explored
  - Visually, there may be a positive relationship
- When a linear regression was run, the model was lacking
  - R value of 0.648
  - As a general rule of thumb, an R value needs to be more extreme than 0.7

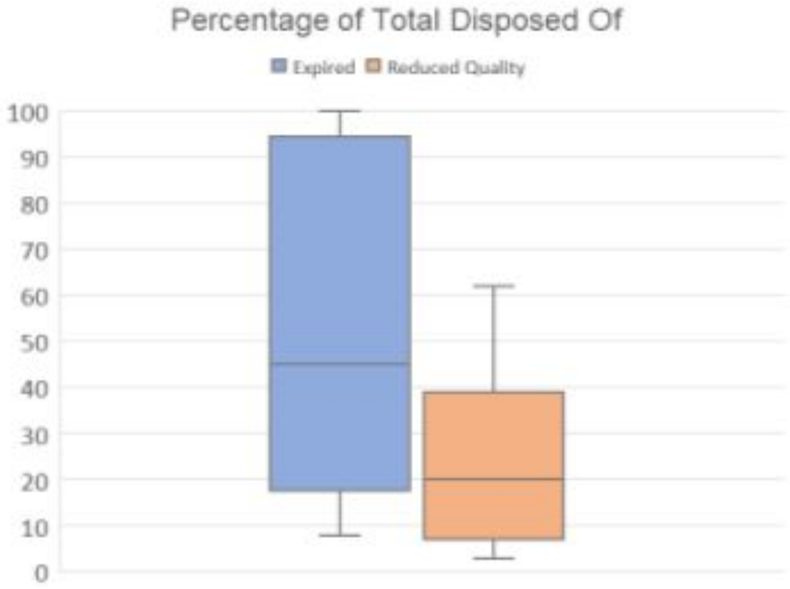




SUMMARY OUTPUT					
Regression Statistics					
Multiple R	0.648765627				
R Square	0.420896839				
Adjusted R Squar	0.403864393				
Standard Error	1.188862541				
Observations	36				
ANOVA					
	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	34.92705214	34.92705214	24.71147369	1.87264E-05
Residual	34	48.0554008	1.413394141		
Total	35	82.98245294			

Final Recommendations

- Two areas where errors occurred
- Boxplots show the percentage of food disposed of in those areas
- Two household rules were implemented:
  - A product would not be replaced if greater than 20% remained
  - Increased efforts to preserve foods



Objectives Covered

Objective 2: Collect and organize data.

- In some cases, data needs to be collected
    - Not being limited by the original data collector’s methods
    - Once data is collected, there are limited ways to increase the rigor of data
    - Be very rigorous in order to ensure accuracy
    - Clearly define data definitions beforehand to ensure clarity

Objectives Covered

Objective 4: Develop alternative strategies based on the data.

- The goal is not only to understand the data, but to draw actionable recommendations
- In MBC 638 - a process improvement plan was implemented.
    - Two areas in the original process were identified where the majority of errors occurred
    - A number of potential alternative strategies were available
- It is important to understand the constraints associated with any alternative strategy
    - Budget
    - Time
    - Ethics

Objectives Covered

Objective 5: Develop a plan of action to implement the business decisions derived from the analyses.

- Stemming from the identification of alternative strategies, a plan of action must be created to implement changes
- In MBC 638 - these rules took place in different stages in the process
  - Actors have to be identified and informed
- The process improvement plan may need to be completed multiple times in an iterative process

Works Cited

2018 Atlanta CYBERATTACK. (2021, January 28). Retrieved February 08, 2021, from https://en.wikipedia.org/wiki/2018_Atlanta_cyberattack

Agrawal, S. (2020, June 03). What is instance-based and model-based learning? [s1e10]. Retrieved February 08, 2021, from https://medium.com/@sanidhyaagrawal08/what-is-instance-based-and-model-based-learning-s1e10-8e68364ae084

Amy Mitchell, J. (2020, August 17). Many Americans SAY made-up news is a critical problem that needs to be fixed. Retrieved February 08, 2021, from http://www.journalism.org/2019/06/05/many-americans-say-made-up-news-is-a-critical-problem-that-needs-to-be-fixed/

Ben.k. (2018, June 11). Text mining preprocess and Naive Bayes Classifier (Python). Retrieved February 08, 2021, from https://medium.com/@baemaek/text-mining-preprocess-and-naive-bayes-classifier-da0000f633b2

CISA insights - Ransomware Outbreak. (n.d.). Retrieved February 08, 2021, from https://www.dhs.gov/blog/2019/08/21/cisa-insights-ransomware-outbreak

Computational complexity of machine learning algorithms. (2018, April 16). Retrieved February 08, 2021, from http://www.thekerneltrip.com/machine/learning/computational-complexity-learning-algorithms/

Cyber daily: Ransomware recovery triggers cultural change in atlanta. (2020). *WSJ Pro.Cyber Security,* Retrieved from https://libezproxy-syr-edu.libezproxy2.syr.edu/login?url=https://www-proquest-com.libezproxy2.syr.edu/trade-journals/cyber-daily-ransomware-recovery-triggers-cultural/docview/2357433761/se-2?accountid=14214

Deere, S. (2018, November 29). Feds: Iranians led CYBERATTACK against Atlanta, other U.S. entities. Retrieved February 08, 2021, from https://www.ajc.com/news/local-govt--politics/feds-iranians-led-cyberattack-against-atlanta-other-entities/xrLAyAwDroBvVGhp9bODyO/

Fruhlinger, J. (2020, February 20). Recent ransomware attacks define the malware's new age. Retrieved February 08, 2021, from
https://www.csoonline.com/article/3212260/recent-ransomware-attacks-define-the-malwares-new-age.html

Works Cited Continued

Gensim.utils.lemmatize()¶. (n.d.). Retrieved February 08, 2021, from https://tedboy.github.io/nlps/generated/generated/gensim.utils.lemmatize.html

Kanani, B. (2020, April 30). Cosine similarity - text similarity metric. Retrieved February 08, 2021, from https://studymachinelearning.com/cosine-similarity-text-similarity-metric/

Palmer, D. (2018, November 28). SamSam ransomware created by iranian hackers, says US DOJ. Retrieved February 08, 2021, from https://www.zdnet.com/article/samsam-ransomware-created-by-iranian-hackers-says-us-doj/

Ransomware hits Atlanta Police dashcam footage. (2018, June 07). Retrieved February 08, 2021, from https://www.bbc.com/news/technology-44397482

Ransomware prevention and response for cisos. (2016, July 14). Retrieved February 08, 2021, from https://www.fbi.gov/file-repository/ransomware-prevention-and-response-for-cisos.pdf/view

Robinson, J. (n.d.). Text mining with r: A tidy approach. Retrieved February 08, 2021, from http://www.tidytextmining.com/tfidf.html

Soumik. (2020, March 09). Goodreads-books. Retrieved February 08, 2021, from https://www.kaggle.com/jealousleopard/goodreadsbooks

Two Iranian men indicted for DEPLOYING ransomware to Extort HOSPITALS, municipalities, and public institutions, causing over $30 million in losses. (2018, November 28). Retrieved February 08, 2021, from https://www.justice.gov/opa/pr/two-iranian-men-indicted-deploying-ransomware-extort-hospitals-municipalities-and-public

What can we learn from Atlanta? (n.d.). Retrieved February 08, 2021, from https://www.govtech.com/security/What-Can-We-Learn-from-Atlanta.html