Applied Data Science Portfolio

(Class Number 37293 – Section M400)

Assignment: Final Paper
Student Name: Ryan Tervo
Due On: June 9, 2023
Submitted On: June 9, 2023

EXECUTIVE SUMMARY:

This paper is an integrated reflection on the knowledge gained from the Syracuse Applied Data Science Master's program. The Syracuse Applied Data Science Master's degree program is a practitioner's program with six learning goals. This paper demonstrates how those learning goals were achieved by synthesizing applicable course projects. The six learning objective are the following:

1. Use programming languages such as R and Python to support the generation of actionable insight
2. Apply visualization and predictive models to help generate actionable insight
3. Communicate insights gained via visualization and analytics to a broad range of audiences (including project sponsors and technical team leads
4. Collect, store, and access data by identifying and leveraging applicable technologies
5. Create actionable insight across a range of contexts (e.g. societal, business, political), using data and the full data science life cycle
6. Apply ethics in the development, use and evaluation of data and predictive models (e.g., fairness, bias, transparency, privacy)

This paper will explore the first three program objectives and how they were met demonstrated by course projects.

INTRODUCTION:

Initially I completed an online Data Analytics certificate program, five courses, from the Air Force Institute of Technology (AFIT) in 2022. I enjoyed the program so much that I decided to pursue a Master's in Applied Data Science. After narrowing the potential school choices down, I carefully considered three. Ultimately I decided that Syracuse University was the best choice and started in the program in September of 2022. After joining Syracuse I was pleasantly surprised that some of my AFIT credits were able to be transferred in, thus, making me a grateful 'Transfer Student'.

During my time at Syracuse I was surprised at the diversity of backgrounds of my fellow students. It seemed like everyone I spoke with had different professional, educational, and Data Science experience. As a result, working with such a diverse set of students was challenging, rewarding, and enhanced my learning.

OBJECTIVE 1: Use programming languages such as R and Python to support the generation of actionable insight

Describe the program learning outcomes

Both Python and R are both open source programming languages that are well suited for building predictive models. Due to the versatility and power, both languages are routinely in the top 20 programming languages available today. For a data scientist these two languages are like likely the two primary languages one will be expected to use.

These languages have the ability to process and clean data to get it suitable for the model to process. Once the model has processed the data then Python and R can then be used to visualize the results. These functions are key to be able to generate insights that are actionable.

Explain how their course projects demonstrated specific program learning outcomes

Course: IST 736 Text Mining.

Project Name: Final Project

Description: The airline industry is a tough business. Fortunately passengers are very vocal is surveys if things don't meet their expectation. Unfortunately processing all of these negative surveys is very time manpower intensive. If this could be automated then it would help the airlines be more responsive and save them resources.

How objective was met: The goal of this final project was to first use sentiment analysis to identify negative flight reviews. Once the negative flight reviews were identified then topic modelling was applied to help identify common issues or complains.

OBJECTIVE 2: Apply visualization and predictive models to help generate actionable insight

Describe the program learning outcomes

This learning objective has three elements to it. First, being able to make predictive models. That means being able to take a data set, prep the data, and successfully run it in a predictive model. Generally this accomplished in the program R or Python. There are many model options but a good rule of thumb is to use the least complex model possible that is still useful. This prevents overfitting of the data and speeds up processing time.

Second, once the model has been demonstrated to be effective the results must be communicated. While statisticians and data science enthusiasts might have the time and inclination to understand the technical details of data science more often than not our audience will not be well versed in statistically significant tests and linearity tests. As such, one of the most effective ways to communicate the results are visually. Care must be taken to make your case while not overstating or misleading the audience.

The third part, arguably the most important from a business perspective, is the ability to generate actionable insight. Now that the data is model is effective and we can demonstrate it visually to the audience the next question is 'so what?' It's important to be able to be able to take the results of the model and translate that into what steps should be taken to improve the business to make it more effective or efficient.

Explain how their course projects demonstrated specific program learning outcomes

Course: IST 719

Project: Final Project, Poster

Description: This course's final project was a large poster. The goal was to tell a story with data.

How objective was met: The goal of this final project was to tell a story with the data. It analyzed 19 years' worth of on time vs. late flight data. Specifically, could a relationship between the airport or airline be identified.

OBJECTIVE 3: Communicate insights gained via visualization and analytics to a broad range of audiences (including project sponsors and technical team leads

Describe the program learning outcomes

Data scientists work with two broad groups, technical group and business group. Successful data scientists must be able to communicate to both audiences. Between the two, generally, the challenge is being able to explain the technical process and results to non-technical persons who need to understand what is going on and be able to evaluate your recommendation.

One of the best communication methods is using accurate and clarify visualizations. If done correctly, the visualization will make a compelling argument that supports a particular action.

Explain how their course projects demonstrated specific program learning outcomes

Course: IST 772

Project: Midterm, project

Description: This course's midterm project was a dataset in which we had to interpret the results. In the scenario we had to interpret the results to two audiences. A technical group and a non technical group.

How objective was met: The write up was written to both a technical and non technical audience. To make the points clear visualizations were incorporated to help make the points and support a course of action.

APPENDIX 1: IST 719 Information Visualization

DOMESTIC AIRLINES: 2004 TO 2022

FLIGHT
DOMESTIC AIRLINES 2022

DESTINATION
SYRACUSE UNIVERSITY

DATA ACCESS CODE

PASSENGER
Tervo, Ryan

PILOT
PROFESSOR KRUDYS, GARY

GATE
IST 719

BOARDING ZONE
Information Visualization

DEPART
28 MAR 2023 9:00 ET

AUDIENCE: Travel Message
Travelers wanting to land on time can use the findings from 2022 to identify which airlines and airports to seek and which to avoid.

Data driven individuals can evaluate how 2022 compared to the prior 18 years.

FROM: Intro
TO: Facts

INTRODUCTION:

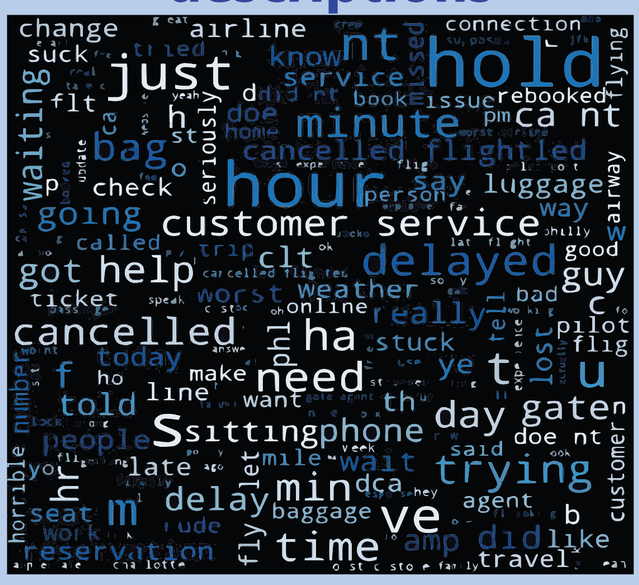
Air travel has become synonomous with frustration. This project examines late flight data from the last 19 years to determine if flying is getting worse or if passenger's expecations have changed.

FACTS AND FIGURES:

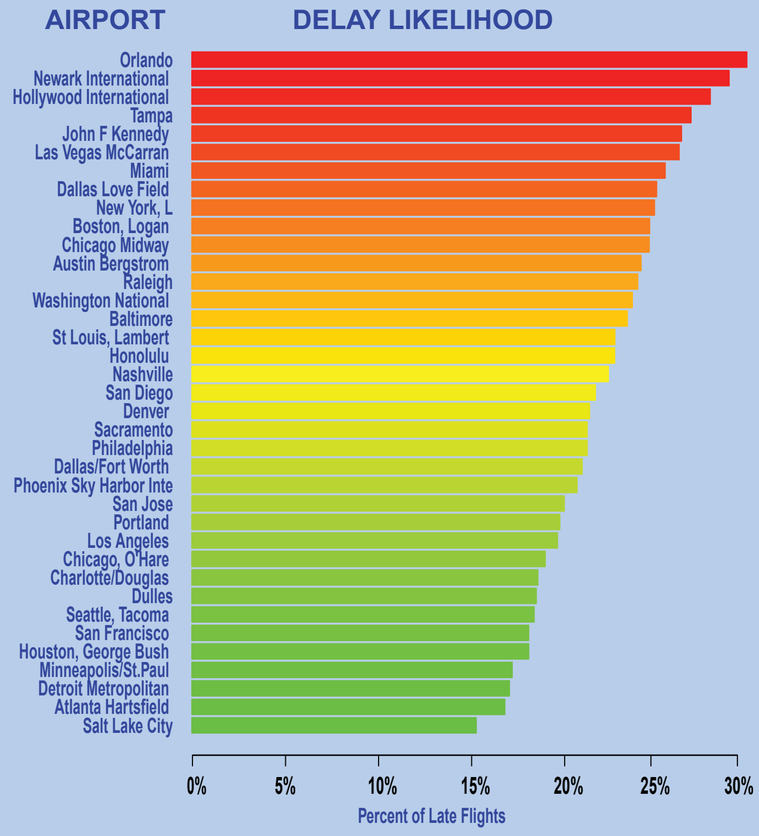
On average, in the U.S. there are 45,000 flights daily with approximately 2.9 million passengers.

* Surveys show one of the top issues are delays.

Top negative flight descriptions

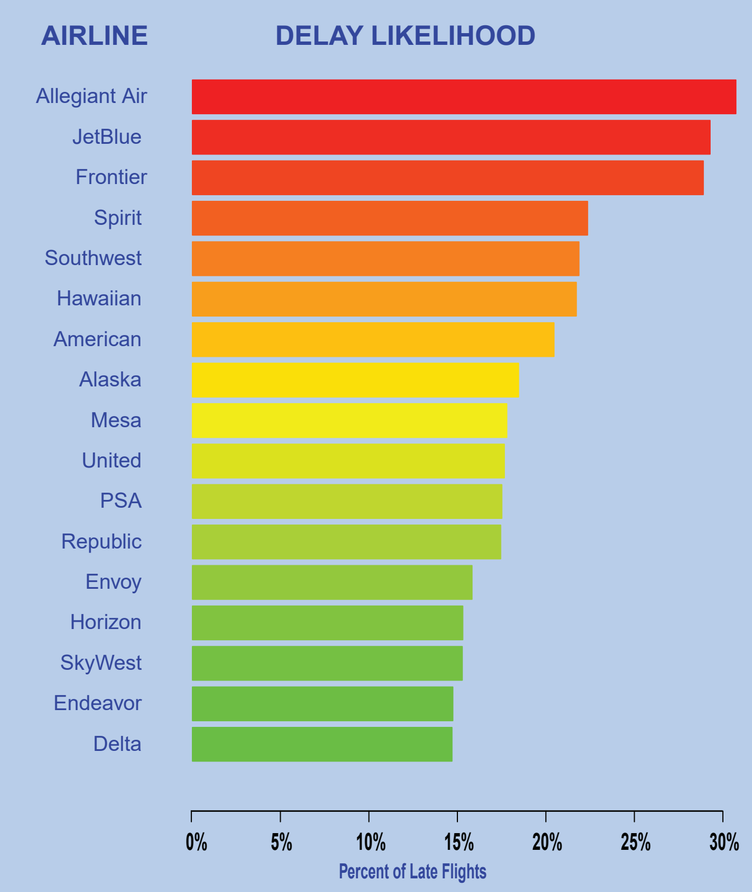


Does the airport make a difference in 2022?



What are the flight trends for the last 19 years?

Does the airline make a difference in 2022?



The trends show that 2022 was a tough year for commercial air travelers. The percentage of late flights is the 3rd greatest in 19 years. Delayed Flights in 2022 on average were nearly 70 mins late which is the second worst rate in 19 years. In addition, 2022 had the highest number of cancellations with the only exception of 2020 which is an outlier due to COVID 19.

In 2022, the airport and airline impacted how often flights are delayed.

* The right airline or airport can decrease your chances of being late by 50%!

About the Data

The data was collected from the Bureau of Transportation Statistics. The raw dataset had 321,110 rows and 21 variables. For analysis purposes rows that reported zero flights were removed. In addition, the delay and cancellation counts were converted into percentages. The data was aggregated by year, airline, and airport for analysis.

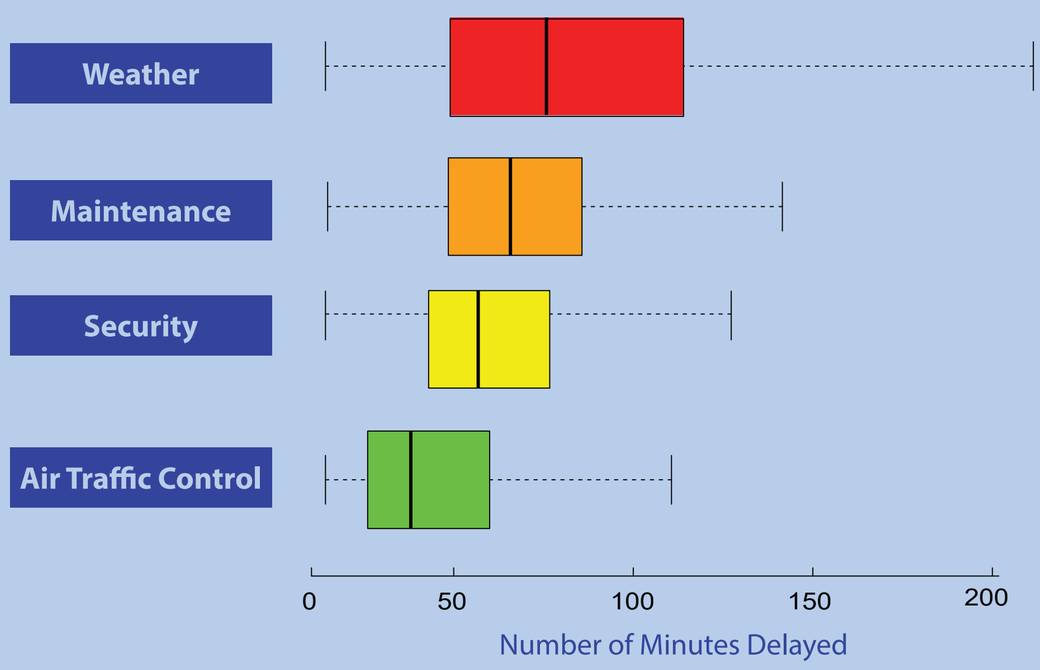
Data Source:

Airline On-Time Statistics and Delays Causes
January 2014 to December 2022
(https://www.transtats.bts.gov/OT_Delay/OT_DelayCause1.asp)

R Packages Used:

dplyr, ggplot2, wordcloud

What are common delay causes and their associated delay times?



APPENDIX 2: IST 736 Text Mining

Sentiment Classification and Topic Modeling for Airline Review Tweets

Meichan Huang, Ryan Tervo
School of Information Studies
Syracuse University
Syracuse, NY, 13210
mhuang01@syr.edu; rtervo@syr.edu

Abstract

This project analyzes a corpus of 14,640 tweets to gauge public sentiment toward six major airlines in the U.S. Through the use of sentiment classification and Latent Dirichlet Allocation (LDA) topic modeling, the project aims to develop an algorithm that can accurately classify sentiment in tweets automatically. Additionally, the project investigates the topic patterns of positive and negative tweets to gain insight into customers' perceptions of airlines. By leveraging machine learning techniques, this study seeks to provide a deeper understanding of the factors that drive sentiment in the airline industry to gain insights into customer preferences and opinions.

1. Introduction

This study seeks to help airlines identify customer frustrations in the travel experience. Two separate, but related, modeling techniques will be used. First, sentiment analysis will be used to identify tweets with negative sentiment.

The second is topic modeling. Topic modeling will be used to identify specific areas that customers are unhappy with. If successful, this will automate an otherwise labor-intensive function and enable airlines to continuously monitor public sentiment towards their travel experience.

A plethora of literature has been published on the sentiment classification system of Twitter data for US airline services analysis using different algorithms and feature engineering (e.g. Rane & Kumar, 2018; Wan & Gao, 2015), and fairly recently, researchers started to tap into the possibility of using topic modeling for online review for airlines (Korfiatis, et al. 2019; Kwon et al., 2021). However, very little research has discussed the potential benefits of combining sentiment analysis and topic modeling in the airline service reviews. Therefore, in this project, we proposed the following three research objectives: (1) What is the best algorithm for predicting sentiment of tweets in the airline industry? (2) What frequent topics are associated with positive sentiment vs. negative sentiment? (3) Do six airline companies receive similar or different topics in travelers' negative comments?

2. Method

2.1 Data Set

The dataset used in this project was obtained from Kaggle.com and comprises 14,640 tweets scraped in February 2015. It focuses on the sentiment of travelers toward six major airlines during that time (Figure Eight, 2015). The data is in csv. format, with 15 columns, containing information regarding the tweets (ID, time, user, re_tweets), the tweet text with all the handles and hashtags, and URL links. The data has also been pre-labeled with sentiments (positive, negative, and neutral) by human annotators, who also categorized negative comments based on reasons such as "late flight" or "rude service".

It is worth noting that the dataset has an unbalanced distribution of data points across the different sentiment classes (Figure. 1). Specifically, there are 9,178 negative comments, 3,099 neutral comments, and 2,363 positive comments.

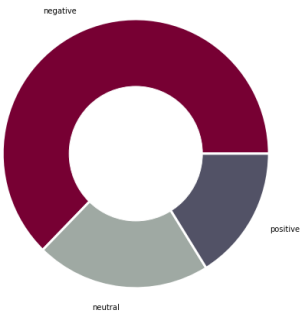


Figure 1. *Number of sentiments in each class*

Moreover, the number of comments received by each airline varies (Figure. 2), with United

Airlines receiving over 3500 reviews and Virgin America receiving approx. 500 reviews.

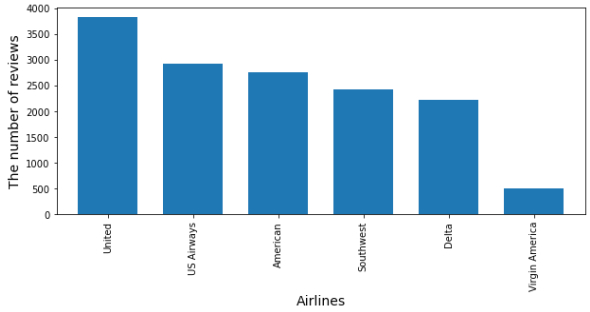


Figure 2. *The number of reviews by airlines*

2.2 Data Preprocessing

After downloading the CSV file, we loaded the dataset into a pandas data frame for preprocessing as our project is focused on sentiment classification and topic modeling, which requires the retention of the text and its respective sentiment labels. Since Twitter data is often inconsistent and has some features that could interfere with the machine's processing of the text, we applied several pre-processing techniques to clean the tweets, such as using Regex to remove potential noise such as URLs, punctuations, and special characters (emojis). Additionally, we applied lowercase conversion and WordNet stemming to remove capitalization and lemmatize the tokens. Stopwords were not removed in this process, considering that they are important in the formation of N-grams. Instead, we utilized the vectorizers' parameter tuning to remove default English stopwords and a list of customized stopwords, including the airline handles (i.e. @southwestair, @usairways).

After pre-processing, we further processed the data by vectorizing it, as both classification algorithms and LDA topic modeling require a document-word matrix as the primary input. We experimented with several vectorizers for the classification tasks, testing and tuning each to determine the best performance. The classifiers were trained using 80% of the data and tested using the remaining 20%. For the LDA modeling, CountVectorizer was used since LDA is based on term count and document count.

2.3 Classification Algorithms

We employed two distinct types of algorithms in order to determine the optimal classification method. Naive Bayes, which is a probabilistic algorithm commonly used for text classification, was one of the algorithms we used. Naive Bayes is known for its simplicity, speed, and ability to operate with relatively little training data. It is also well-suited for handling high-dimensional data and can handle missing values. Additionally, we employed Support Vector Machines (SVMs), which are effective for handling high-dimensional data, especially when the data is linearly separable. SVMs are capable of handling both categorical and continuous data and are ideal for small to medium-sized datasets.

Specifically, six model configurations were investigated to determine the best method to identify negative tweets. The Multinomial Naive Bayes (MNB) and Support Vector Machine (SVM) algorithms were used in combination with three vectorizers. To fit these two algorithms, the data were vectorized using two different Term-frequency Counts (CountVectorizer) settings and a term frequency-inverse document frequency (TF-IDF) Vectorizer with parameters fine-tuned to various degrees, i.e. min_df, max_df, stopwords removal, and so on.

2.3.1 Classification Vectorizer Configuration

The classification models attempt to correctly categorize tweets as either 'positive', 'negative', and 'neutral'. The word vectorizer's parameters were modified to maximize the model's performance. In order to evaluate the model's performance the model's accuracy, precision, recall, F-1 scores, and the top 10 positive and negative words were used. Using trial and error the following settings appeared to provide the best overall performance in terms of performance and top 10 negative and positive words.

Table 1. *Vectorizer 1 Settings*

Parameters	Setting
encoding	'latin-1'
binary	False
min_df	3
max_df	1500
ngram_range	(1, 5)

Table 2. *Vectorizer 2 Settings*

Parameters	Setting
encoding	'latin-1'
binary	False
min_df	2
max_df	1500
ngram_range	(1, 3)

Table 3 *Vectorizer 3 Settings*

Parameters	Setting
encoding	'latin-1'
binary	False
min_df	5
smooth_idf	True
sublinear_tf	True
ngram_range	(1, 4)
max_features	2000

2.3.2 Classification Model Interpretation

The sentiment classification algorithms are designed to predict the tweet's sentiment as either 'negative', 'positive', or 'neutral'. This is accomplished by first vectorizing the tweets. The model is then trained to associate those vectors (features) with their respective labels. The trained models are then evaluated using test data to determine how well they can predict the correct label.

In this use case both the overall model accuracy and specifically how well the model can predict 'negative' tweets are considered the most important. Due to the importance of being able to identify one particular label ('negative') the specific recall, precision, and F1 score will be used to evaluate the model. These scores allow us to compare multiple model's performance.

2.4 Topic Modeling Toolkit

In order to conduct the evaluation of the topics in customers' opinions on airlines, this project used Latent Dirichlet Allocation (LDA) (Blei et al., 2003) for topic discovery. LDA is a statistical modeling technique that helps identify the hidden topics present in a given corpus or dataset. It posits that each document is a mixture of a small number of topics and that each word's creation is attributable to one of the document's topics. The project applied Scikit-learn's built-in LDA module (Pedregosa *et al.*, 2011) for text processing, document classification, clustering, and topic modeling.

2.4.1 Topic Modeling Interpretation

We conducted topic modeling using the LDA model using sklearn.de To determine the optimal number of topics that best encompassed the information in the corpus, we tested several different numbers of topics, including K20, K10, K5, K4, K3, and K2, each with a max_iter of 10. We computed the log-likelihood scores and perplexity scores for each model to evaluate its performance. Additionally, we conducted model comparisons using GridSearch to determine the best estimator for the number of components.

2.4.2 Human Judgment on Topic Coherence

The outputs of the LDA topic modeling were used to examine the differences in themes or topics. To assess topic cohesion and meaningfulness, we conducted a human judgment evaluation through the composition of the top 10 words in the topic clustering outputs, as well as the comparison of LDAVis outputs for each topic modeling, following the suggestion of Chang et al. (2013). Our evaluation served as a triangulation method for the statistical modeling outputs in sklearn, such as log-likelihood and perplexity scores. As Kapadia (2019) suggested, "optimizing for perplexity may not yield human interpretable topics", as is the case we found out later in the topic modeling process.

3 Results

In the following section, we discuss our final results with the classification and topic modeling. For more specific results, please see the attached html output of our codes.

3.1 Classification

The six model configurations were evaluated using a ten-fold cross validation. Given the focus on 'negative' tweets, we chose to report the 10-cv accuracy scores, recall, precision, and F1 scores

that were based on the 'negative' category only. In Table 4 the models were evaluated using the data labeled as either 'negative', 'positive', and 'neutral'.

The six models were evaluated a second time in which the category labels were modified to be either 'negative' and 'notNegative'. The 'notNegative' category was the combination of 'positive' and 'neutral' tweets. Table 5 shows the results in which the precision, recall, and F1 scores that are based on the 'negative' category only.

Table 4 *Model Results using 3 Categories*

Model	Vectorizer	Accuracy	Precision	Recall	F1-Score
MNB	Count 1	75%	80%	88%	84%
MNB	Count 2	75%	79%	88%	84%
MNB	TF-IDF	74%	74%	96%	83%
SVM	Count 1	73%	82%	83%	83%
SVM	Count 2	74%	82%	84%	83%
SVM	TF-IDF	76%	81%	89%	85%

Table 5. *Model Results using 2 Categories*

Model	Vectorizer	Accuracy	Precision	Recall	F1-Score
MNB	Count 1	79%	81%	87%	84%
MNB	Count 2	80%	82%	87%	84%
MNB	TF-IDF	79%	78%	92%	85%
SVM	Count 1	78%	82%	82%	82%
SVM	Count 2	78%	83%	83%	83%
SVM	TF-IDF	80%	82%	86%	84%

All of the model configurations, using the data categorized using two or three labels, performed better than the trivial model. The best accuracy was achieved using the SVM model with the TF-IDF vectorizer which improved the accuracy of the trivial model by 13.1%. The trivial (no information) model would predict every tweet 'negative' and achieve an accuracy of only 62.9%.

In evaluating the SVM model using the TF-IDF vectorizer the F1 score was 85% (original categories) and 84% (two categories). The F1 score is of particular interest because it is the harmonic mean between the recall and precision scores.

3.2 Topic modeling

This report presents the results of topic modeling on positive and negative comments received by airline companies. Our findings show that positive tweets about airlines tend to be more general in nature, while negative comments can be classified into three major topics.

We also analyzed whether companies received negative feedback for similar or different reasons. Our results indicate that while the words associated with different topics varied across different airlines, there were some common trends in the negative aspects of airlines that are frequently reported.

For more details about the log likelihood scores and perplexity for all the modeling in this project, please refer to the attached html output of the coding and outputs.

3.2.1 Topics Distribution for Positive Tweets

The analysis indicates that travelers tend to provide a general overview of their positive experiences. Figure x illustrates that keywords associated with positive experiences include smooth flights and positive feedback regarding crew/attendant services.

Figure 3. *WordCloud on all positive comments*

Although the model with n_components = 2, when compared to other numbers of components, has the highest log likelihood (-93538.20085373441) and the lowest perplexity score (723.0571290935178), and indicated the better model fit for the data theoretically (citation needed), 3 topic clusters actually worked the best with topic coherence and topic interpretability, when human judgments were factored into for model evaluation. Based on the output of top 10 keywords in topic modeling, when the corpus was reduced to 3 topics, it encompasses a distinct set of topics without obvious overlapping.

Table 6. *Top 10 keywords with 3 topics in positive comments*

Topic	Top 10 words
1	flight, great, thanks, crew, time, gate, response, plane, home, good
2	service, thanks, customer, customer service, help, good, flight, today, thx, amp

3	thank, thanks, just, love, guy, airline, best, fly, yes, flying

Upon reviewing the LDAvis results, it becomes apparent that analyzing the top 30 words associated with each topic provides greater insight into the underlying themes. In general, the feedback from customers is positive regarding their experiences with airline customer service and flight experiences. Within the flight experience category, the quality of attendant/crew service and a seamless travel experience stand out as key themes within the dataset. It appears that positive feedback from travelers is primarily focused on the services provided by airlines and a smooth overall experience, while negative feedback is less centered on these aspects.

3.2.2 Topic Distribution for Negative Tweets

When examining the topics associated with negative tweets, three major themes emerge: (1) punctuality issues such as cancelled or delayed flights; (2) poor customer service characterized by rude agents and slow response times; and (3) problematic baggage handling leading to lost or damaged luggage. Even if the top 10 words for each topic varied slightly in different companies, the general themes seemed to align with the general negative comments received across the airline companies, as shown in Table 7:

Table 7. *Top 10 keywords with 3 topics in negative comments*

Topic	Top 10 words
1	time, hold, hour, agent, airline, aa, minute, worst, phone, wait
2	hour, plane, delayed, bag, late, delay, waiting, gate, hr, time
3	cancelled, customer, service, flightled, help, cancelled flightled, customer service, day, need, amp

When deciding which was the best estimator, n_component = 3, again, although this model did not have the highest log likelihood and lowest perplexity score compared to other models, there was no overlapping amongst topics, based on the intertopic distance map using LDAvis (see Figure. 4). All topics are semantic meaningful.

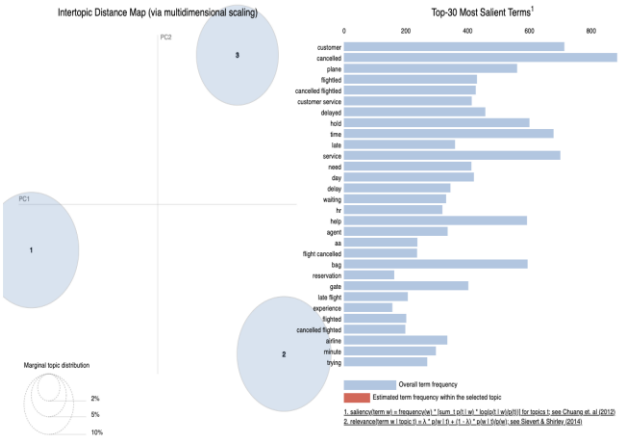


Figure 4. *LADvis output for negative comments*

In addition, the topic modeling effectively aligns with the overall themes identified in the human annotation of the original dataset's negative reviews, apart from the tweets that even human annotaters cannot tell, see Figure 5 below:

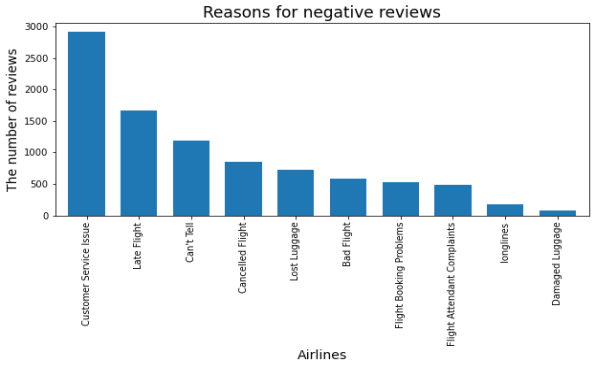


Figure 5. *Reasons for negative reviews assigned by human annotators in the original dataset*

3.2.3 Most Frequent Topics for Negative Tweets by Airline

Despite the variations in comments and word choices received by each airline, it appears that these differences do not significantly affect the overall topic composition of negative feedback expressed by travelers. In general, the identified topics appear to center on the same three problematic areas of airline services: (1) customer service; (2) flight punctuality issues such as cancellations or delays; and (3) luggage handling. However, some minor variations were also noted, including issues with check-in services and email communication with customers, as shown in Table 8.

Table 8. *Comparison of Topic Keywords across different airlines (n_component = 3)*

Topics	United	US Airway	AA	Southwest	Delta	VA
1	service, customer, customer service, thanks, like, just, help, know, need, did	flight, hold, hour, time, service, customer, help, delayed, need, minute	flight, cancelled, did, ca, seat, help, just, flighted, cancelled flighted, hold	flight, cancelled, cancelled flightled, flightled, hold, flight cancelled, hr, flight cancelled flightled, time, cancelled flightled flight	just, need, help, right, fleek, fleet, fleet fleek, ca, blue, work	flight, cancelled, check, customer, jfk, virgin, today, airline, vx, ca
2	flight, bag, plane, cancelled, hour, gate, time, amp, seat, late	day, bag, customer, service, luggage, phone, people, doe, say, customer service	service, hour, customer, agent, customer service, gate, time, bag, phone, hold	hour, hold, flight, phone, help, ca, online, trying, hold hour, just	flight, delayed, time, service, customer, amp, cancelled, customer service, hour, min	email, website, site, doe, day, checkin, contact, iphone, problem, applied
3	flight, delay, hour, airline, delayed, ua, time, luggage, crew, worst	flight, cancelled, plane, flightled, cancelled flightled, delay, mile, sitting, flight cancelled, hour	flight, cancelled, flightled, cancelled flightled, hour, aa, late, dfw, flight cancelled, delayed	customer, service, customer service, bag, hour, flight, plane, gate, late, seat	flight, delay, jfk, plane, hour, bag, just, gate, hr, time	eat, help, flight, trying, time, luggage, ticket, just, hour, book

The following Figures 6. - 11. visually display the word clouds generated from negative tweets received by each airline. These graphics provide compelling evidence that certain topic themes are commonly shared across all airlines when it comes to the negative experiences reported by passengers.

Figure 6. *WordCloud on United's negative comments*

Figure 7. *WordCloud on US Airway's negative comments*

Figure 8. *WordCloud on American Airlines' negative comments*

Figure 9. *WordCloud on Southwest' negative comments*

Figure 10. *WordCloud on Delta's negative comments*

Figure 11. *WordCloud on Virgin America's negative comments*

5. Discussion and Conclusions

The objective of this project was to assess the public sentiment towards six major airlines in the U.S. using a corpus of 14,640 tweets and to understand the topics that generate these sentiments. To achieve this goal, sentiment classification and Latent Dirichlet Allocation (LDA) topic modeling were employed to develop

an algorithm capable of accurately classifying sentiment in tweets, and exploring topics the topic patterns of both positive and negative tweets, providing valuable insights into customers' perceptions of airlines.

The sentiment analysis showed that SVM using the TF-IDF vectorizer performed best. The model achieved an accuracy of 76% using the three label categories. Evaluating the data using only two labels, 'negative' and 'notNegative', the accuracy improved to 80%. In addition, with the emphasis on 'negative' tweets the F1 score was used to determine how well the model is identifying 'negative' tweets. Achieving an F1 score of 84% shows that the model is performing very well in identifying 'negative' tweets.

The results of the topic modeling suggest that when discussing their positive experiences with airlines, travelers tend to provide less specific information. This result is not unexpected since people tend to express their compliments less frequently and may not always provide specific reasons for positive experiences unless they are particularly outstanding, such as exceptional service or an extremely friendly flight attendant.

However, negative comments align well with common real-life issues we experience with airlines, including flight cancellations or delays, dissatisfaction with customer service during the booking process, poor service provided by flight attendants, and lost or damaged luggage.The fact that all airlines received similar complaints in these areas suggests that airlines can use sentiment analysis and topic modeling of online reviews as a tool to identify common customer preferences and concerns. By scraping all online reviews, airlines can gain valuable insights into what their customers like and dislike about the services they provide, helping them to improve the customer experience and stay competitive in the industry.

The results also has meaningful applications in the airline industry. The results should be considered as a successful attempt in trying to automate the classification process of public sentiment and gain insights into these sentiments. That is, first, classifying large volumes of online reviews can be a time-consuming task. Applying classification algorithms can automate the process, saving time and increasing efficiency. Applying classification algorithms can ensure a consistent approach to analyzing online reviews, reducing the risk of bias and errors that may arise from manual analysis.

In addition, measuring the quality of services is a challenging task because different customers may have varying preferences. However, by extracting negative comments from airline passengers' online reviews and performing topic modeling, companies can gain an understanding in areas with shared negative customer experience, where they need to improve their services and enhance the customer experience, leading to more satisfied customers and increased revenue. The findings suggest that using quality features extracted from tweets can help predict variations in passenger preferences and competition, indicating their potential usefulness in evaluating airline service quality.

However current project also has its limitations, which will constrain the generation of the above results. First, the performance of the models might be far from perfect. For classification models, there are indeed better models suggested in previous literature, e.g. neural network models (RNN) and AdaBoost (Rane & Kumar, 2018), a technique in which multiple classifiers were combined; however, due to our limitation, we were not able to explore the possibility these models can provide.

Second, regarding the topic modeling, even if the current project has used both statistic outputs and human judgment for topic modeling, the performance of the topic modeling still needs further examination, since the determination of the number of topics can be arbitrary and depends on the researchers' ability to interpret the resulting topic solution.

Lastly, sklearn libraries provided different LDA modeling tools, with gensim being one of them. We tried the Gensim module and modeled the positive and negative comments with n_components = (5, 4, 3) respectively and the results slightly different but eventually, since we are more familiar with sklearn, we opted to use sklearn for our analysis. One advantage of using gensim is that Gensim's LDA has more built in functionality and applications for the LDA model such as Topic Coherence Pipeline or Dynamic Topic Modeling, in addition, the built-in parameter of coherence scores would help adding more statistical measures that human judgment or log likelihood scores.

References

[1] Blei, D. et al., (2003) Latent dirichlet allocation. *Journal of machine learning research*,(4-5):993-1022.

[2] Chang, J., Gerrish, S., Wang, C., Boyd-Graber, J., & Blei, D. (2009). Reading tea leaves: How humans interpret topic models. *Advances in neural information processing systems*, 22.

[3] Wan, Y., & Gao, Q. (2015, November). An ensemble sentiment classification system of twitter data for airline services analysis. In *2015 IEEE international conference on data mining workshop (ICDMW)* (pp. 1318-1325). IEEE.

[4] Kapadia, S. (2019). Evaluate Topic Models: Latent Dirichlet Allocation (LDA): A step-by-step guide to building interpretable topic models. Retrieved from: https://towardsdatascience.com/evaluate-topic-model-in-python-latent-dirichlet-allocation-lda-7d57484bb5d0

[5] Korfiatis, N., Stamolampros, P., Kourouthanassis, P., & Sagiadinos, V. (2019). Measuring service quality from unstructured data: A topic modeling application on airline passengers' online reviews. *Expert Systems with Applications*, *116*, 472-486.

[6] Kwon, H. J., Ban, H. J., Jun, J. K., & Kim, H. S. (2021). Topic modeling and sentiment analysis of online review for airlines. *Information*, 12(2), 78.

[7] Rane, A., & Kumar, A. (2018, July). Sentiment classification system of twitter data for US airline service analysis. *In 2018 IEEE 42nd Annual Computer Software and Applications Conference (COMPSAC)* (Vol. 1, pp. 769-773). IEEE.

[8] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., & Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *The Journal of machine Learning research*, 12, 2825-2830.

APPENDIX 3: IST 772 Quantitative Reasoning for Data Science

Quantitative Reasoning
(IST 772-M401: 37284)

Assignment: Midterm
Student Name: Ryan Tervo
Due On: February 28, 2023
Submitted On: February 25, 2023

PART 0: Assignment Instructions:

General Instructions:
This is an honor system exam that is open book and open notes. You may consult any of the feedback I have provided to you on homework or practice exams. You may not confer or collaborate with any human besides me.

Problem Scenario:
A startup company has developed an inexpensive and environmentally friendly biofilm to remove dissolved solids in water treatment plants. TDS is an abbreviation that refers to "total dissolved solids" and is measured in parts per million (PPM). Lower TDS is better – it means the water is cleaner. The startup conducts a comparison of batches of dirty water with and without their new treatment. The control group contains batches of water processed using industry standard mechanical filtering methods. The treatment group contains batches of water filtered with the biofilm. The research (alternative) hypothesis is that the mean TDS in the treatment group will be lower than the mean TDS in the control group. Specially calibrated, highly sensitive devices are used to measure TDS, so each control and treatment batch costs a lot of money to run.

The company will not release the raw data because they consider it a trade secret, but they have provided the following statistical outputs for you. Your job is to produce a report that will guide their biologists and investors on the next steps for this project. As such, the company wants you to evaluate the research hypothesis and write an interpretation of it that can be understood by non-statisticians. Here's the output that they provided to you. You can feel free to cut and paste any of the graphics that appear below into your report, as appropriate for the audience:

Data Received
The data received will be placed in the answer section of the report.

QUESTION 1: (1 point)

What are the lower bound and upper bounds of the (frequentist) 95% confidence interval of the mean difference?

The 95% confidence interval of the mean difference is [-12.35, 14.30].

The lower and upper bounds of the (frequentist) 95% confidence interval of the mean difference is provided using program 'R', 't.test'. The 't.test' compares the means between the control group and the treatment group as seen below in Provided Data Results 1.

```
> t.test(x=testDF[1],y=testDF[2])

     Welch Two Sample t-test

data:  testDF[1] and testDF[2]
t = 0.14925, df = 31.048, p-value = 0.8823
alternative hypothesis: true difference in means is not
equal to 0
95 percent confidence interval:
 -12.35187  14.30250
sample estimates:
mean of x mean of y
 32.05410  31.07879
```

Provided Data Results 1: Welch Two Sample T- Test

QUESTION 2: (1 point)

What is the point estimate of the mean difference?

The point estimate of the mean difference is 0.97.

The point estimate of the mean difference is the difference of the point estimates for each mean.
The point estimates for the means are approximately 32.05 and 31.08.
The difference of these two means is 32.05 – 31.08 = 0.97

QUESTION 3: (1 point)

Report the outcome of the null hypothesis significance test on the difference of means. Make sure to state the null hypothesis.

Due to the results of the 't-test' we "fail to reject the null hypothesis."

The null hypothesis is that there is no difference between the two means. The alternative hypothesis is that there is a difference between the two population means.

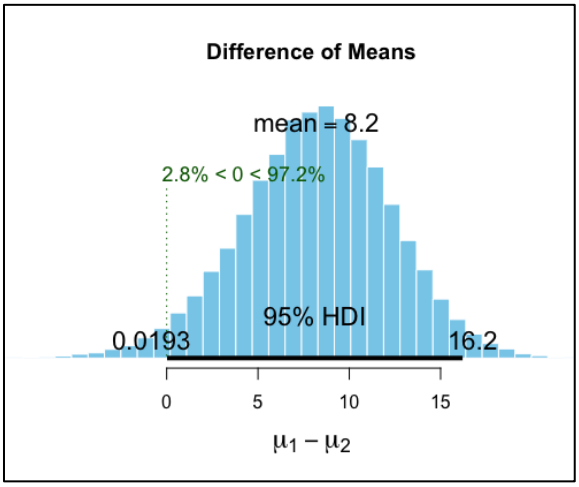
Using the 't.test' results, the confidence interval is [-12.35, 14.30] which spans zero and the p-value is 0.88 which is much greater than 0.05. For both of these reasons there is not enough evidence to support rejecting the null hypothesis, thus we fail to reject the null hypothesis. This is not to say we accept the null hypothesis just we do not reject it.

QUESTION 4: (1 point)

Report the lower and upper bounds of the 95% Highest Density Interval for the difference of means.

The 95% Highest Density Interval for the difference between means is [0.0193, 16.2].

As seen in the Provided Data Results 2, the 95% HDI is identified on the bottom of the graph in the dark line on the x axis. It spans from 0.0193 to 16.2.



Provided Data Results 2: BEST mcmc Plot of Mean Differences.

QUESTION 5: (1 point)

Report the percentages of values in the posterior distribution of mean differences that are above zero and below zero.

The posterior distribution mean difference above zero is: 97.2%
The posterior distribution mean difference below zero is: 2.8%

These results are identified on the Data Results #2 graph in green font. In the green font it show that 2.8% of the results are less than zero and that 97.2% is greater than zero. This is strong evidence that the mean difference is in fact greater than zero.

QUESTION 6: (5 point)

Write a 1-2 paragraph technical report. The technical report should contain the detailed information that it would be important for other statisticians to know about the data, about the analytical results, about any anomalies you observed, and about how any such anomalies may have affected the reported results. You can cut and paste any of the graphics included above, as long as you provide a 2-3 sentence explanation of what the graphic means.

The results of the statistical tests shows a conflicting results between the 't-test' (Frequentist) and the 'BESTmcmc' (Bayesian). The 't-test' had a confidence interval of [-12.35, 14.30] and a p-value of 0.88. With those results we failed to reject the null hypothesis, the null hypothesis being that the two means are the same. To be clear, this is not evidence that the means are the same only that we do not have evidence to believe that they are not the same.

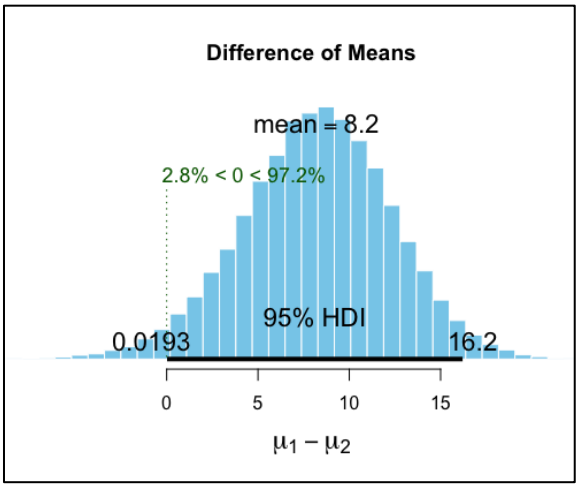
```
> t.test(x=testDF[1],y=testDF[2])

     Welch Two Sample t-test

data:  testDF[1] and testDF[2]
t = 0.14925, df = 31.048, p-value = 0.8823
alternative hypothesis: true difference in means is not
equal to 0
95 percent confidence interval:
 -12.35187  14.30250
sample estimates:
mean of x mean of y
 32.05410  31.07879
```

Provided Data Results 1 (again): Welch Two Sample T- Test

When examining the Bayesian BESTmcmc results it shows a stark contrast. The Bayesian analysis showed that there was likely 97.2% chance that the means are not the same and that likely difference between means is close to 8.2. In addition, the Highest Density Interval was between [0.0193, 16.2] which does not span zero. All of this strongly indicates that there is a difference between the means.

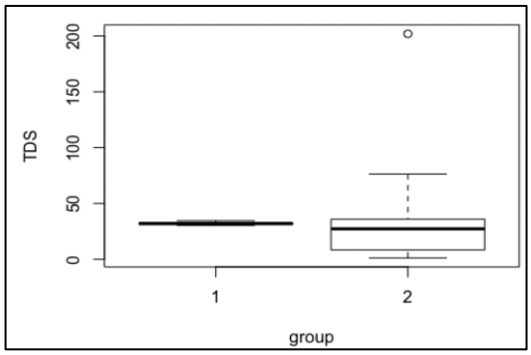


Provided Data Results 2 (again): BEST mcmc Plot of Mean Differences.

In light of the BESTmcmc results the underlying data was examined to try to help identify why the ‘t-test’ failed to reject then null. In looking at the basic structure of the sample data and box plots, it appears there are many differences between the two sample groups that might indicate what occurred to lead to these results. See Provided Data Results 3 for reference.

First, the range between the control and the treatment is vastly different. For the treatment its range is from 30.3 to 34.5, which is a difference of slightly more than 4. In contrast the treatment had a range of 1.2 to 201.9, this is a difference of over 200. The box plots show these differences very well. The box plot also shows the treatment group max value being an extreme and rare outlier.

```
> str(testDF)
'data.frame':  32 obs. of  2 variables:
 $ Control  : num  32.8 33.2 30.3 33.2 31.9 ...
 $ Treatment: num  35.1 16.5 31.6 11.8 201.9 ...
> summary(testDF)
    Control        Treatment
 Min.   :30.27   Min.   :  1.178
 1st Qu.:31.25   1st Qu.:  8.501
 Median :31.90   Median : 27.259
 Mean   :32.05   Mean   : 31.079
 3rd Qu.:32.81   3rd Qu.: 35.533
 Max.   :34.53   Max.   :201.908
```



Provided Data Results 3: Basic Structure Data and Control/Treatment Box Plots

In reviewing the structure of the data it also shows that there are significant differences between the control and treatment first quartile. The control has a first quartile being 31.25 which means the first 25% of the values range from 30.27 to 31.25 while for the treatment has 25% of its values ranging from 1.18 to 8.5. This is a substantial decrease, compared to the control group.

Given that the two means are 32.05 and 31.08 it is not surprising that the ‘t-test’ failed to reject the null. However, if the extreme outlier (201.9) from the treatment group was removed then the treatment mean would have been calculated to be 25.6. This shows that the outlier significantly impacted the control mean result.

Given the weight of the evidence (the ‘t-test’, BESTmcmc, and underying data) it is appears that the control method’s strength is that it can consistently remove TDS to less than 34.5 ppm but can only decrease it down to about 30 ppm at best. The new treatment’s strength is that it has the ability to decrease TDS to less than 35.5 ppm about 75% of the time and greatly reduce the TDS below the standard method about 25% of the time. However, the new treatment appears to be significantly less effective than the control group 25% of the time in which the TDS was greater than 35.5 ppm. This means that while the new method can outperform the current standard method it does not do so consistently.

QUESTION 7: (5 point)

Write a 1-2 paragraph report of the results of your analysis for presentation to the company's biologists and investors. This report should be in plain language, interpretable by non-statisticians. Make sure to integrate the Bayesian evidence, the frequentist confidence interval, and the results of the null hypothesis significance test.

The biologists and investors need to decide what the startup should do next: The essential question they want to answer is whether or not the biofilm shows promise as an alternative to traditional filtering techniques. Use the results of these statistical analysis to provide them with guidance.

The results of the statistical tests of Null Hypothesis Significants Test (NHST), frequentist confidence interval, and the Bayesian evidence showed very interesting results. From the samples provided it appears that NHST and frequentist confidence interval did not have enough evidence to show if there was or was not a difference between the two methods. However, the Bayesian (BESTmcmc) results did show that there was a high likelihood that the new biofilm method can outperform the standard method.

In reviewing the underlying data it appears that the standard method consistently brought the TDS to the low to mid 30's every single time. The biofilm on the other hand sometime was significantly below 30 and sometimes significantly above 35. This means that while the biofilm does have the ability to decrease the TDS from the water lower than the current standard method it does not do so consistently. This is likely why the NHST and frequentist confidence interval was inconclusive while the BESTmcmc supports the biofilm being able to outperform the standard method.

Final Recommendations:

1) The data provided supports the belief that the biofilm has the ability to outperform the standard method but its lack of consistent performance is problematic. Given how inconsistent the biofilms results were it is recommended that additional research and development (R&D) be conducted to further refine the biofilm until its performance range is more consistent. In all, the biofilm has a lot of potential but is not ready for prime time… at least not ready yet!

2) Also, it is also recommended that the sample that provided a reading of over 200 ppm be examined to ensure that this was an accurate result. An outlier this large is a bit suspect. It might be an accurate reading or could be a data entry mistake or a result of an error in the application process. With that said, it would be prudent to investigate this data point to determine its legitimacy.

It was a pleasure working with you and your new company. We wish you the best luck in the future. Please contact us again if we can ever support your research data analysis needs.