OMSA 6740 – COMPUTATIONAL DATA ANALYSIS/ MACHINE LEARNING I

Final Report - Fall 2023

Name: Brandon Jones

Project Title: PCA-Squared; or, an Inquiry into whether Google Maps Review Text can Predict Denomination: A Survey of Machine Learning Techniques

Problem Statement:

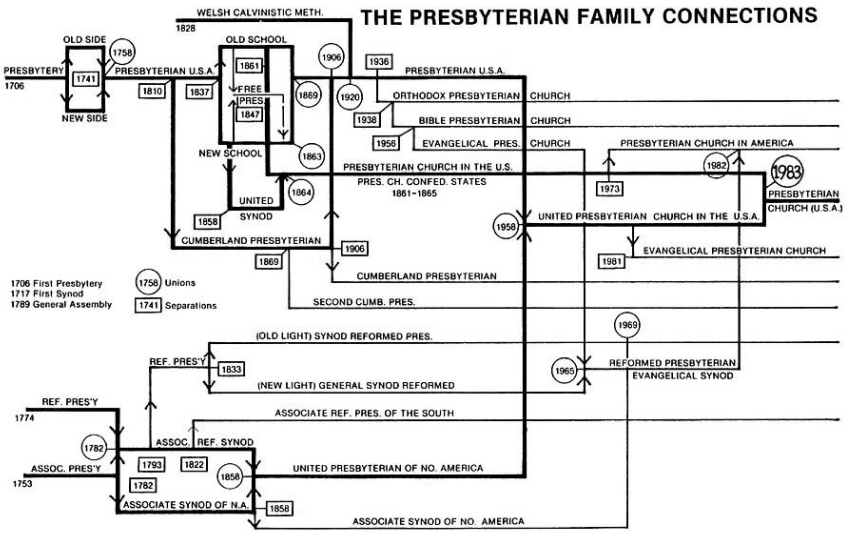
A religiously observant person may be interested in finding a house of worship while on vacation, business trip, or otherwise away from home. In those cases, sometimes the person would be looking on, say, Google Maps. Now, supposing our religious person is a Presbyterian, we assume he is going to look for a Presbyterian church.

But are all Presbyterian churches the same? Indeed, they are not. In fact, there are (sometimes large!) differences among Presbyterian churches, some significant, others not so significant; some having to do with ecclesiastical history, some having to do with doctrine, some having to do with particular worship practices, etc.

As a rule, one of the high-level labels associated with churches is their denomination. This is true for Presbyterian churches, too. There are many Presbyterian denominations, and even if we ignore the "micro" denominations, there are easily a dozen and a half or so which make up the bulk of the Presbyterians in North America. For example, the largest is the Presbyterian Church (USA), known as the PC(USA), with around 9,000 congregations and over 1 million active members, while the second largest, the Presbyterian Church in America (PCA) has around 380k members, with many smaller denominations with 10 – 30k members, such as the Orthodox Presbyterian Church (OPC).

Denominations form for a variety of reasons, but without getting into the particulars, which are beyond the scope of this investigation, we can fairly state that the primary reason is that "birds of a feather flock together".

Table 1: A Summary of Major Presbyterian Denominational Splits



It may be that for the discerning church-goer, one or another of these particular denominations would be desirable, while others would be unacceptable. Often, churches do not include their denomination as part of their name, moreover, a typical review may not also include this information, and so on a quick Google Maps search it could be difficult to directly ascertain to which denomination a particular church belonged.

Yet, we believe there may be implicit information about the church contained in the reviews themselves, such that without explicitly stating the denomination, we wonder if we can detect those similar themes, priorities, and emphases which are theoretically common to churches in the same denomination.

Data Source:

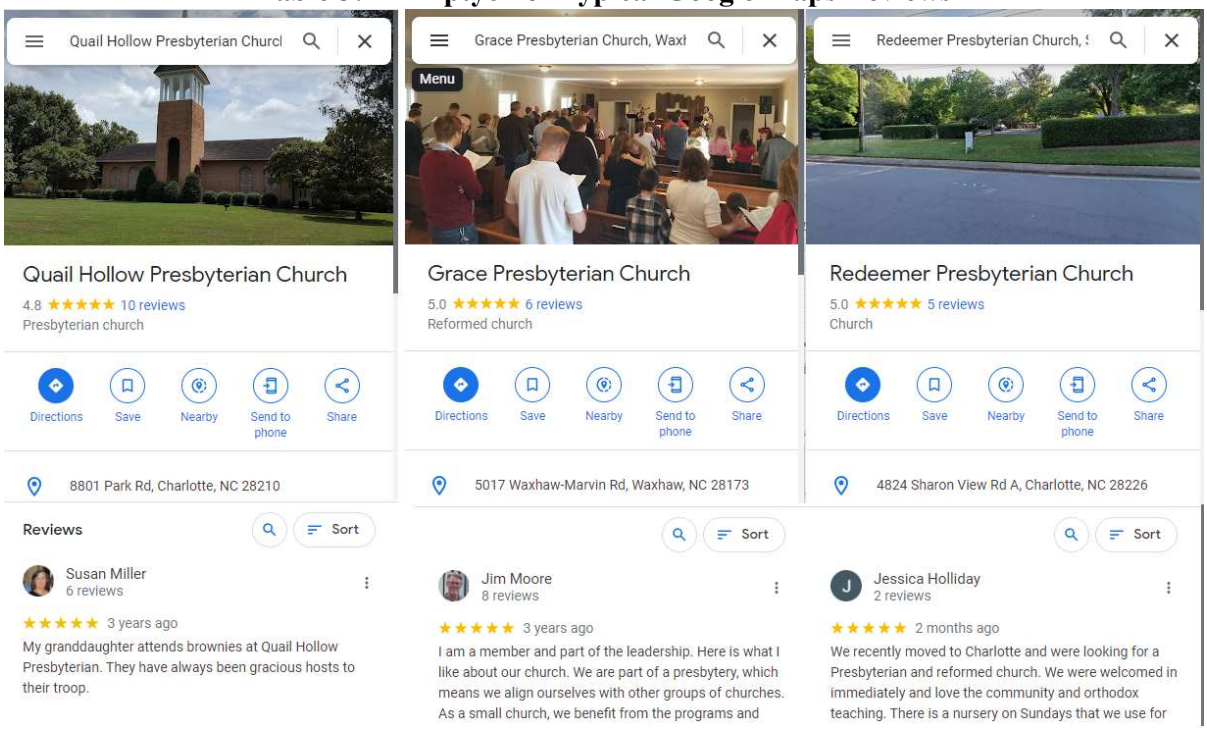
Although Google does have an API for review collection, by using Selenium WebDriver[i] with python scripts, we can semi-manually scrape the review text for whatever Google Maps location we would like for free.

For the purposes of this analysis, we used 100 churches from each of the PCA, PC(USA) and OPC. For the PCA and PC(USA) those 100 congregations were chosen from the Southeastern United States of America – Georgia, Florida, North Carolina, South Carolina, Alabama, Mississippi, and Tennessee. The OPC's smaller size required a larger sampling area.

For this analysis, we needed not only to merely identify 100 congregations, but 100 congregations with reviews with text (not bare star rankings). To ensure enough data for analysis, we set a soft target of approximately 5 text-including reviews as being suitably large.

Typical reviews are shown below, and vary in length and scope.

Table 3: A Triptych of Typical Google Maps Reviews



High Level Methodology:

Google Maps allows users to leave reviews of all kinds of establishments, including churches[ii]. These are free-form text. Once scraped into suitable data structures, we will analyze using a **Bag of Words** approach.

Bag of Words is a Natural Language Processing (NLP) approach in which the words that comprise a set of documents are counted for frequency.

Table 2: Bag of Words Example

	it	was	the	best	of	times	worst	expect	prepare	for
It was the best of times it was the worst of times	2	2	2	1	2	2	1	0	0	0
Expect the best prepare for the worst	0	0	2	1	0	0	1	1	1	1

In actuality, many of the most common words in a corpus of text are going to be uninteresting (‘the’, ‘and’, ‘in’, ‘of’, ‘or’, ‘for’,), and are typically excluded from analysis as “stop words”. In our analysis, a typical list of English stop words was used, as well as ('place', 'translated', 'great', 'people', 'church'), which were common enough to be unhelpful (‘translated’ was from the text of the google-supplied translations of reviews left in Spanish or Korean).

Another consideration when working with text data is whether or not to engage in lemmatizing and stemming, which are two ways of standardizing text to aid in analysis. Stemming aims to get at the root of a word, and lemmatizing tries to regularize tense, number, and inflection. Our analysis was done with and without these approaches, and we achieved better results without either.

Source of Potential Methodological Concern:

1. Congregations were manually selected. Although we did not intend to bias the selection in anyway, as the process required inspecting each potential congregation for analysis, it is possible included congregations are not completely representative in some way. Due to the time and effort involved in manually copying hundreds of URLs, it was not possible at this time to collect a larger sample from which to run analysis on a sub-sample.

2. Although the PCA and PC(USA) churches were only from the Southeast, the OPC churches span the USA. However, we do not believe there would be any notable geographical bias.

3. To find congregations large enough to have enough reviews, we focused on the largest 8-10 urban areas of each state, and using each denomination’s handy congregation locator, manually inspected the candidate congregations Google Maps pages. Therefore, insofar as these results are of any use at all, they may only be applicable for urban/sub-urban congregations. There is good reason to consider that rural congregations may differ from urban/sub-urban congregations**.**

Denomination Scope

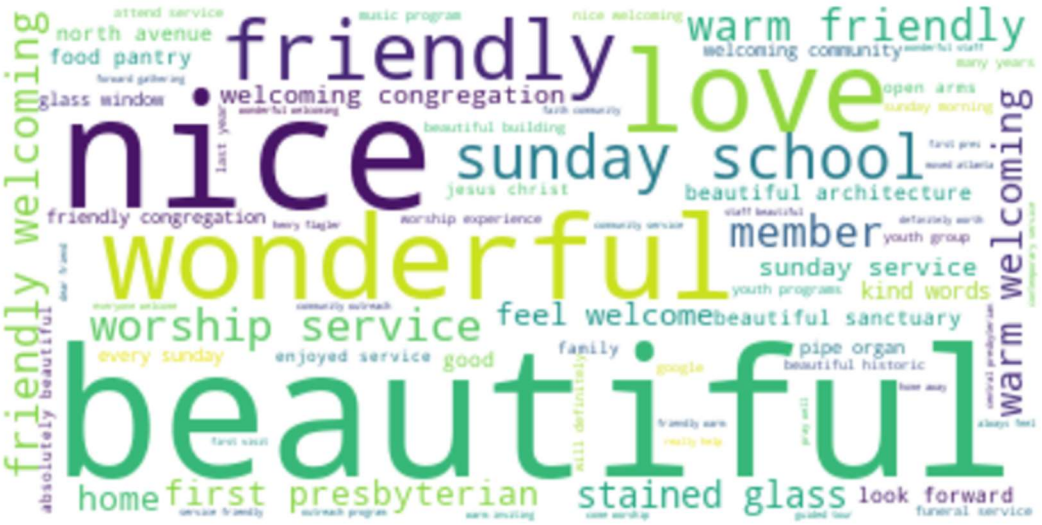
Although there are many Presbyterian denominations, for this exercise we restricted ourselves to 3: PC(USA), PCA, and OPC. If the reader will indulge us not defining “conservative” and “liberal” in this context, it is assumed a priori that among those three denominations, the PC(USA) would be the most liberal, the OPC the most conservative, and the PCA somewhere in between, though more akin to the OPC than to the PC(USA). Ideally, these differences will be different enough to be measurable in the Google Maps reviews.

Preliminary Data Exploration

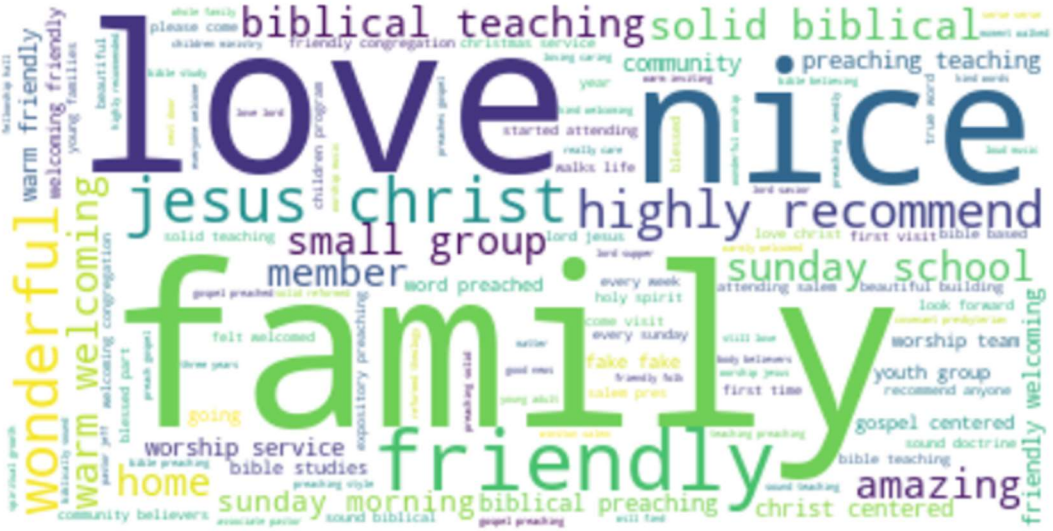
Before diving into the Machine Learning algorithms, we thought it could be interesting as a first step to inspect the relative frequency and importance of particular words and 2-grams (two-word phrases) in the review text of the denominations taken individually.

While WordClouds are not ML techniques strictly speaking, they are sometimes interesting and suggestive. Though possible to create WordClouds from a Bag of Words matrix (sometimes called a Document Term Matrix), that approach only looks at words individually. It seemed more interesting to look at the WordClouds when 2-grams were included as well. See below for the WordClouds for PC(USA), PCA, and OPC.

Graph 3: PC(USA) Word Cloud of Google Maps Review Text



Graph 4: PCA Word Cloud of Google Maps Review Text



Graph 5: OPC Word Cloud of Google Maps Review Text

It does appear from a cursory glance that there may be a difference in the general content of Google Maps review text according to Presbyterian denomination. Without going into too much detail, we note the following in passing:

1. “love” appears prominently in all three WordClouds
2. “family”, “jesus christ”, “biblical preaching/teaching” appear in PCA and OPC, not in PC(USA)
3. “nice” and “friendly” appear in the PC(USA) and PCA, but not in OPC
4. “beautiful” appears in PC(USA) only, and seems to often be connected with the actual church building, as opposed to the church as a congregation

It would be hard to come to any particular conclusions from a mere WordCloud. After all, are churches in the OPC not “nice” and “friendly”? Maybe. But, at least for our purposes it suggests we may be able to find enough variance in the text to come up with a good predictor.

Evaluation and Final Results

All reviews for a particular congregation were concatenated into a large review. Therefore, we used 1 column of text per congregation. Using sklearn’s **CountVectorizer**, our Bag of Words matrix contained 300 rows (pertaining to the 300 congregations) and 5547 columns (pertaining to the 5547 unique words used across all reviews for all congregations). This matrix is sparse and of somewhat high dimension, which presents somewhat of a challenge for analysis.

Naïve Bayes

It is appropriate to start with a Bayesian analysis as Bayes was himself a Presbyterian minister. In any case, we first used the Bag of Words directly, and achieve 70% accuracy. This is a very good result for us, since random chance would have about 33% accuracy.

Code Snippet 6

```
# Naive Bayes
from sklearn.naive_bayes import BernoulliNB

classifier = BernoulliNB()
classifier.fit(X_train, y_train.to_numpy().astype(np.int64))

# Predict Class
y_pred = classifier.predict(X_test)

# Accuracy
from sklearn.metrics import accuracy_score
accuracy = accuracy_score(y_test.to_numpy().astype(np.int64), y_pred)
print('Accuracy: %2.2f %%' % (100. *  accuracy))
```

```
Accuracy: 70.83 %
```

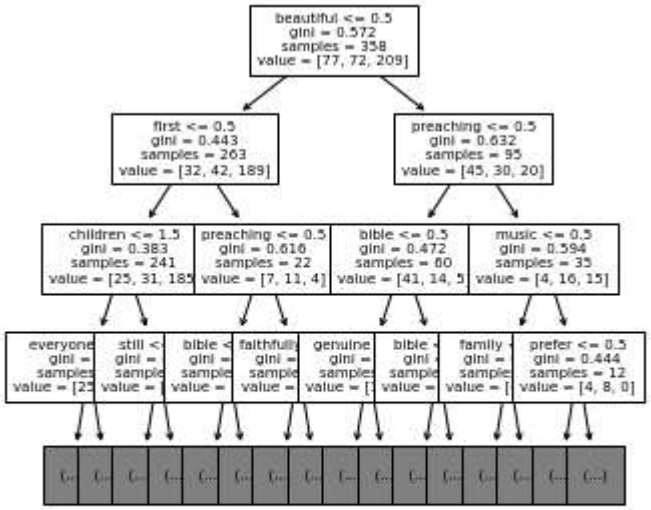
We can also extract the feature log probabilities from the classifier, and find the most predictive words for the 3 classes, though there is significant overlap.

```
Most Predictive Words for PC(USA):
 ['friendly', 'beautiful', 'wonderful', 'community', 'love', 'welcoming', 'service', 'nice', 'pastor', 'good', 'family', 'congr
egation', 'worship', 'god', 'always', 'presbyterian', 'warm', 'home', 'sunday', 'go']


Most Predictive Words for PCA):
 ['friendly', 'love', 'god', 'worship', 'pastor', 'family', 'preaching', 'teaching', 'welcoming', 'congregation', 'wonderful',
'nice', 'community', 'word', 'christ', 'home', 'gospel', 'service', 'good', 'sunday']


Most Predictive Words for OPC:
 ['god', 'preaching', 'worship', 'friendly', 'word', 'congregation', 'love', 'christ', 'good', 'pastor', 'welcoming', 'bible',
'biblical', 'fellowship', 'wonderful', 'teaching', 'reformed', 'service', 'gospel', 'warm']
```

Decision Tree

We can also take a Decision Tree approach to classify these reviews into their particular denominations. A Decision Tree over the Bag of Words will, at each node, chose a best classifier to split the tree. In our case, “beautiful” was the always the first node selected. This makes sense given its prominence in the Word Cloud. Interestingly, “first” was a second line selection, which has to do with the fact that of the congregations, 25 “First Presbyterian” churches were PC(USA), while only 7 “First Presbyterian” churches were PCA or OPC. Therefore a “First Presbyterian” is more likely to be PC(USA). We achieve 66% accuracy with this approach.

Graph 7: Decision Tree Features

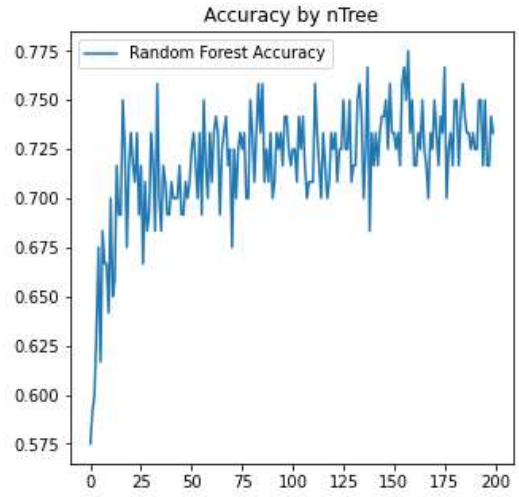



```
<Figure size 90432x90432 with 0 Axes>

Accuracy: 66.67 %
```

Random Forest

With a Random Forest, we create individual decision trees, and combine them to form a classifier which is better together than any one of them is by itself. In our case, a RF technique improves classification to around 73-75%, as long as we use around 100 trees.

Graph 8: Random Forest Accuracy by Number of Trees Used




```
Accuracy: 73.33 %
```

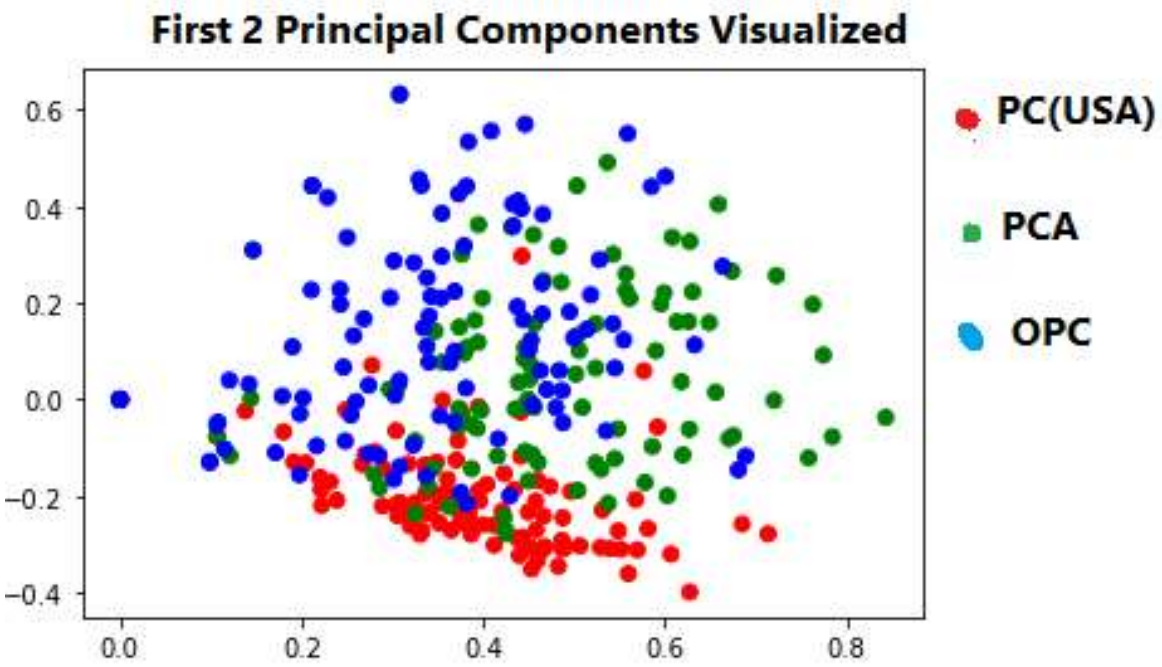
TF-IDF and Principal Component Analysis Analysis

Another way of looking at the most important words in a corpus is term frequency–inverse document frequency (TF-IDF), which seeks to balance mere term frequency in a document with how common it is across all documents. For example, “the” is a common word in all documents, so the fact that is appears often in a particular document does not mean much. But with TF-IDF we can rank words in terms of a more suitable importance than mere frequency.

Additionally, to this TF-IDF-transformed data, we also performed Principal Component Analysis (PCA). PCA is a dimensionality reduction technique which preserves to most information given a number of principal components. In the Bag of Words matrix, there are likely way more dimensions than needed to capture the variance across denominations, so we can simplify processing by taking only the top 2 principal components. Additionally, taking 2 principal components will aid us in visualizing as well.

On inspection, we do not see clearly separable categories – there is a lot of overlap. Yet, we do see distinct regions. The PC(USA) reviews are more or less all together. The PCA and OPC, overlap together much more than they overlap with the PC(USA).

Graph 9: PCA of Google Maps Review Text



While the principal components do not related directly to any one word, and so makes analysis somewhat more difficult, for interest’s sake, we can pull the feature names which contribute the most to the first two principal components. This shows us the “most important” words; that is, which words contain the most variance across the observations, and therefore are more predictive by their value (or absence).

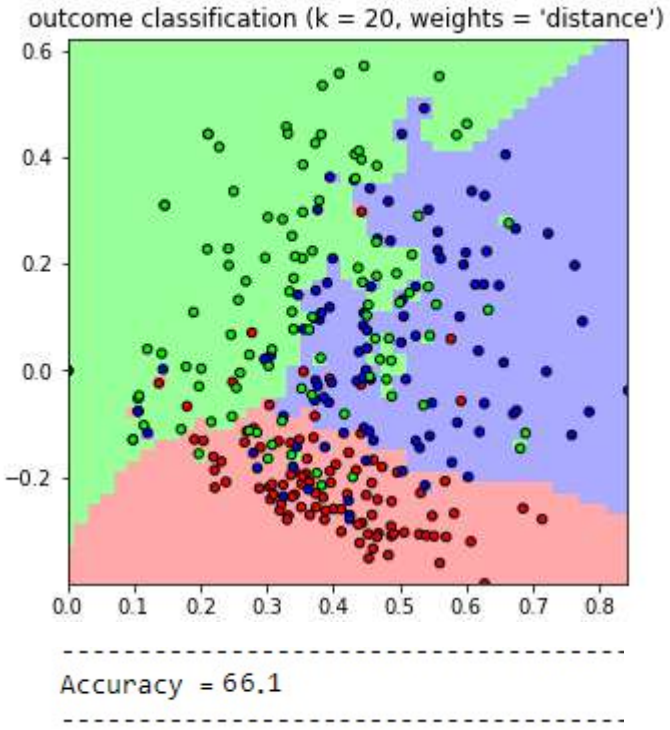
Table 10: Most Important Words to First Two Principal Components

1	bible
2	beautiful
3	fellowship
4	sunday
5	warm
6	gospel
7	home
8	biblical
9	years
10	will
11	one
12	loving
13	lord
14	jesus
15	presbyterian
16	come
17	reformed
18	visit
19	excellent
20	solid

KNN

K-Nearest Neighbors classifies test data points based on the majority vote of the k-nearest neighbors. For this, we are required to split the data in training and test groups. The 'k' parameter can be adjusted to find the best fit. In our case, we achieve the best results with k=20.

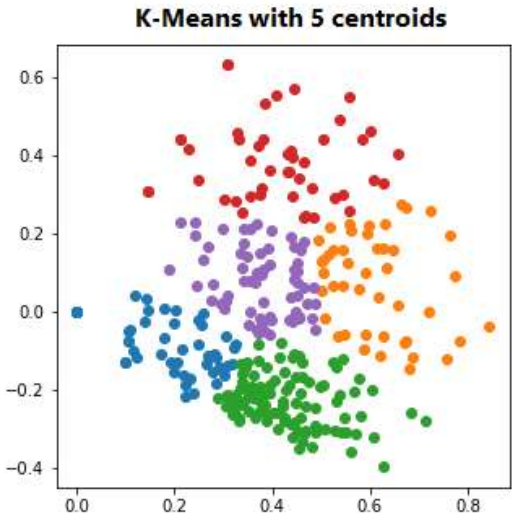
Chart 11: KNN Decision Boundaries



K-Means

While not a classification technique, we were interested to see how a K-Means approach would work. In this case 5 centroids achieves somewhat reasonable performance.

Chart 13: K-Means Clusters



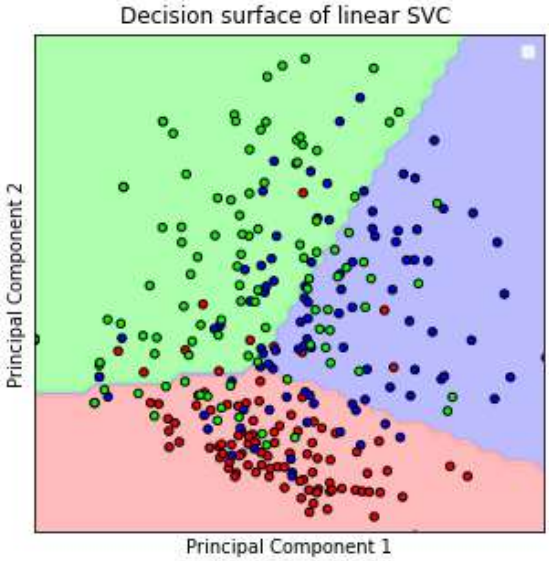

```
-------------------------------------------------
Accuracy = 64.078
-------------------------------------------------
```

SVM

Using support vector machine, we are able to split the data points space into distinct areas. In this case mis-classification in inevitable. However, performance was solid at 74% accuracy.

Chart 13: SVM Decision Boundaries

```
Kernel:  linear
Principal Components:  2
Num Support Vectors:  [55 71 68]
Accuracy = 74.194
```



Conclusion

Overall, we are pleased with the results, especially as we had no idea when proposed if it was even going to work or not!

We find that, yes, it is possible to differentiate these three Presbyterian denominations according to their Google Maps Reviews. Given 3 classes, random chance would have accuracy at 33%, whereas our two best models, Random Forest and SVM topped out at nearly 75%.

The Word Cloud and lists of important words/features makes sense. PC(USA) churches are much more likely to be older congregations, and in older buildings. PCA and OPC congregations are much less likely to have notable ("beautiful") architecture. Similarly, both the PCA and OPC left the PC(USA) over issues pertaining to Biblical fidelity, so it is also not surprising that the PCA and OPC have "biblical" words and ideas as identifying markers in their reviews, especially viz-a-vis the PC(USA).

Methodologically, a logistic regression, or other method, which could provide a probability of classification would be an interesting approach. This could be extended to quantify how good of a fit a congregation's reviews are with the rest of the denomination. However, we aren't sure there is much actual use for this, apart from general curiosity. Originally this was under proposal for this report, but resources did not permit that extension, given the high level of manual work in pulling the reviews.

This analysis could be continued at a more granular level. For example, a given Presbyterian denomination is composed of presbyteries. It would be interesting to see if presbyteries (which are generally regionally defined) differ significantly from one another even in the same denomination.

Table 14: Results

Table of Results

	Accuracy
Naïve Bayes	71%
Decision Tree	67%
Random Forest	73%
KNN	66%
SVM	74%
K-Means	64%

[i] Methodology adapted from https://www.linkedin.com/pulse/webscrape-google-map-reviews-using-selenium-python-choy-siew-wearn/?trk=pulse-article_more-articles_related-content-card

[ii] For example, https://www.google.com/maps/place/Grace+Presbyterian+Church/@34.9295855,-80.763835,17z/data=!4m7!3m6!1s0x885429e83374b19b:0xa0d1729282f8d78a!8m2!3d34.9295394!4d-80.7617648!9m1!1b1