Stat 5353 - Fall 2017
Preliminary Report

Dalton Cole
Adam Harter
Samuel Richter

November 14, 2017

Our analysis of variance, Table 1, shows that there a statistically significant difference in the means of the treatment combinations, so we can say that the experimental variables tested have an effect on the response variable. Our next course of action will be to do further analysis on the data to determine what, if any, individual or interaction effects the experimental variables have on the output, using effects tests to determine significance of effects and comparisons of means using Tukeys test to determine which means are significantly different from others. From this data, we will conclude the best combination of treatment effects to maximize the output, in order to find which distance metric and number of clusters will result in the greatest number of correct answers output by the machine learning algorithm.

The data used to run the statistical analysis is shown in Table 2.

Table 1: ANOVA Table

Source	d.f.	SS	MS	F-ratio
Model	8	16247.444	2030.93	44.4189
Error	9	411.500	45.72	Prob > F
Combination Total	17	16658.944		<.0001

Table 2: Experimental Data

Distance Metric	Number of Clusters	Number Correct
Euclidean	2	180
Euclidean	3	139
Euclidean	5	105
Cosine	2	184
Cosine	3	149
Cosine	5	104
Jaccard	2	176
Jaccard	3	151
Jaccard	5	108
Euclidean	2	180
Euclidean	3	138
Euclidean	5	99
Cosine	2	186
Cosine	3	137
Cosine	5	101
Jaccard	2	174
Jaccard	3	136
Jaccard	5	128