Take home final- CSC424
Sarah Cummings

1. Dataset information:

a. For this project, I am using the file titled BurgersOriginal.sav that I saw used in a youtube video about clustering. The data file is attached with my submission.

b. This dataset contains nutritional values for sandwiches from fast food chains. There are 15 variables:

1. Meat- A nominal value for which type meat is in the sandwich. There are 7 categories: beef, chicken, fish, ham, other, turkey, and veggie
2. Type-A nominal value for what kind of sandwich the case is classified as: burger, crossover, just meat, nuggets, sandwich, wrap.
3. Sandwich- A string value containing the name of the sandwich.
4. Calories- A numeric value for the number of calories in the sandwich.
5. TotalFat_g- A numeric value for the total fat in the sandwich, measured in grams.
6. Sodium_mg- A numeric value for the sodium in the sandwich, measured in milligrams.
7. CaloriesFromFat- A numeric value for the number of calories from fat in the sandwich.
8. SaturatedFat_g- A numeric value for the saturated fat in the sandwich, measured in grams.
9. TransFat_g- A numeric value for the trans fat in the sandwich, measured in grams.
10. Cholesterol_mg- A numeric value for the amount of cholesterol in the sandwich, measured in milligrams.
11. Carbohydrates_g- A numeric value for the carbohydrates in the sandwich, measured in grams.
12. Fiber_g- A numeric value for the fiber in the sandwich, measured in grams.
13. Sugar_g- A numeric value for the sugar in the sandwich, measured in grams.
14. Protein_g- A numeric value for the protein in the sandwich, measured in grams.
15. Restaurant- A nominal value containing the name of the restaurant the sandwich is from. There are seven different restaurants in total: Arby's, Burger King, Chik-Fil-A, Dairy Queen, McDonalds, Sonic and Wendy's.

c. There are 190 cases in this data set

d. We are missing the Calories from fat value for all of our 24 Wendy's sandwiches. To account for this, I will not be using that variable in my analysis, because it would affect the cluster membership for the Wendy's sandwiches.

2. Research question:

Suppose someone on a diet would like to find implicit categorization of sandwiches to help them make smarter choices while eating out at fast food restaurants. They would like to form groupings of sandwiches that could be thought of as providing different levels of nutrition. These sandwiches might be thought of as being different levels of "healthy".

3. Techniques to be used:

For this unlabeled data, I will use k-means clustering to form groupings. I'm using this method because our data is unlabeled, and because it efficiently finds inherent groups for data. I will first standardize the variables since they are on different scales, and then I will cluster with k=3. I will also use LDA so that I can visualize my groupings on a 2-D axis, like we did for one of our homeworks.

4. Tests that the technique can be used:
For this study, my main goal is to use clustering on the available data. Other than standardizing the data, I do not need to perform any prior procedures for the clustering to form groups. I will have Box's M plot and Wilk's lambda tests with my LDA, but since I am just using LDA for visualization, I am not concerned with the results. Unfortunately our Box's M test came out significant, but I wasn't planning on using the LDA scores much anyways. Our Wilk's Lambda, however, was significant.

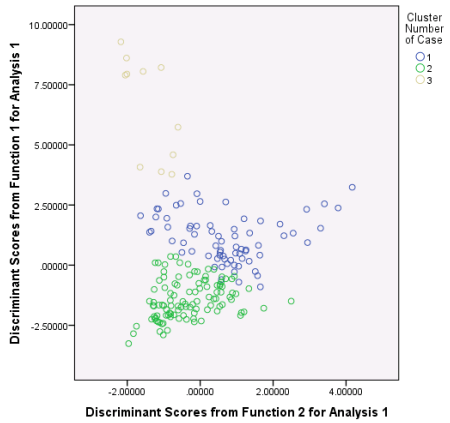
5. See my take-home output file which contains the results from all of my tests. I will copy and discuss some of those results here.

Final Cluster Centers

	Cluster		
	1	2	3
Zscore(Calories)	.61296	-.64546	2.38086
Zscore: Total Fat (g)	.46991	-.58637	2.72404
Zscore: Sodium (mg)	.67620	-.53294	.86788
Zscore: Saturated Fat (g)	.30056	-.50903	3.05772
Zscore: Trans Fat (g)	.17939	-.40291	2.79797
Zscore: Cholesterol (mg)	.37166	-.54326	2.93496
Zscore: Carbohydrates (g)	.64448	-.42579	.02070
Zscore: Fiber (g)	.53318	-.32062	-.29355
Zscore: Sugars (g)	.44681	-.33915	.44586
Zscore: Protein (g)	.51312	-.58924	2.47328

```
GRAPH
  /SCATTERPLOT(BIVAR)=Dis2_1 WITH Dis1_1 BY QCL_2
  /MISSING=LISTWISE.
```

Graph



ANOVA

	Cluster		Error			
	Mean Square	df	Mean Square	df	F	Sig.
Zscore(Calories)	67.012	2	.294	187	227.943	.000
Zscore: Total Fat (g)	67.218	2	.292	187	230.368	.000
Zscore: Sodium (mg)	35.712	2	.629	187	56.799	.000
Zscore: Saturated Fat (g)	68.622	2	.277	187	247.940	.000
Zscore: Trans Fat (g)	52.966	2	.444	187	119.237	.000
Zscore: Cholesterol (mg)	68.218	2	.281	187	242.685	.000
Zscore: Carbohydrates (g)	24.538	2	.748	187	32.793	.000
Zscore: Fiber (g)	16.117	2	.838	187	19.225	.000
Zscore: Sugars (g)	14.392	2	.857	187	16.798	.000
Zscore: Protein (g)	61.740	2	.350	187	176.211	.000

The F tests should be used only for descriptive purposes because the clusters have been chosen to maximize the differences among cases in different clusters. The observed significance levels are not corrected for this and thus cannot be interpreted as tests of the hypothesis that the cluster means are equal.

As seen in the ANOVA table above, all of our z-score variables were significant and useful to our model. They all have p values= 0.00. Looking at the final cluster centers, we see that cluster two has all negative values. This is interesting, and implies that the sandwiches in this cluster are lower calorie, lower fat, lower everything, really. I was hoping to find some sandwiches that were lower calorie lower fat, but higher in protein and fiber. To me, that would imply the sandwiches are good for you. However, I am not a nutritionist. Cluster three contains the sandwiches that are the worst for you, with significantly higher coefficients for calories, total fat, saturated fat,

and trans fat. Cluster one thus contains the sandwiches that are somewhat in the middle— not too bad for you, but not too good for you either.

6. Validate results:
Our model is validated in that it is has significant F statistics for all variables in the ANOVA displayed above.

7. Report results and answer question:
Based on our test, I would recommend that someone who is watching their weight order sandwiches that belong in cluster two. They should also avoid sandwiches from cluster three at all costs. Below, I have presented some tables to help understand which sandwiches were placed in each cluster. As you can see, all of the sandwiches in cluster three were beef hamburgers. This isn't to say all hamburgers are bad for you, as 39 beef based sandwiches were in cluster two. However, if unsure of what to order, it might be best to stick with chicken, 54 of the 88 chicken sandwiches were in group two.

As far as picking a restaurant, Arby's and Chik-Fil-A are the safest bet since they both don't have any cluster three sandwiches. Sonic is the riskiest restaurant choice, with 6 sandwiches in group 3.

Cluster Number of Case * Meat Crosstabulation

Count

		Meat							Total
		Beef	Chicken	Fish	Ham	Other	Turkey	Veggie	
Cluster Number of Case	1	30	34	1	2	0	4	0	71
	2	39	54	4	3	2	2	4	108
	3	11	0	0	0	0	0	0	11
Total		80	88	5	5	2	6	4	190

Cluster Number of Case * Restaurant Crosstabulation

Count

		Restaurant							Total
		Arby's	Burger King	Chick-fil-a	Dairy Queen	McDonald's	Sonic	Wendy's	
Cluster Number of Case	1	14	9	3	7	14	16	8	71
	2	15	22	7	10	24	15	15	108
	3	0	1	0	1	1	6	2	11
Total		29	32	10	18	39	37	25	190

Finally, I went back and created a table of means for each variable for each cluster. I used the unstandardized variables because they present a revalue understanding in units people have seen before. This table again confirms cluster three as containing the least healthy sandwiches, with a mean number of calories at 1017 and mean grams of total fat at 68.0

Report

Cluster Number of Case		Calories	Total Fat (g)	Sodium (mg)	Carbohydrates (g)	Fiber (g)	Sugars (g)	Protein (g)	Saturated Fat (g)	Trans Fat (g)	Cholesterol (mg)
1	Mean	651.97	34.04	1496.62	52.94	3.49	8.55	33.58	10.310	.873	91.76
	N	71	71	71	71	71	71	71	71	71	71
	Std. Deviation	128.781	8.873	420.815	16.973	1.698	4.167	5.879	3.9590	.7593	21.976
2	Mean	391.94	18.13	970.65	36.19	2.14	5.38	21.02	5.222	.319	51.67
	N	108	108	108	108	108	108	108	108	108	108
	Std. Deviation	89.989	6.310	298.458	11.462	1.329	3.580	6.696	2.2921	.5136	20.336
3	Mean	1017.27	68.00	1580.00	43.18	2.18	8.55	55.91	27.636	3.364	204.09
	N	11	11	11	11	11	11	11	11	11	11
	Std. Deviation	178.835	16.162	179.053	2.136	.603	1.368	11.344	6.2012	.8090	47.844
Total	Mean	525.32	26.96	1202.47	42.86	2.65	6.75	27.73	8.421	.703	75.47
	N	190	190	190	190	190	190	190	190	190	190
	Std. Deviation	206.630	15.065	434.996	15.649	1.586	4.033	11.393	6.2842	.9510	43.822