Lab 5 Clustering – Formatted Output and Answers

1. I am building a predictive model using clustering to group Las Vegas buffets according to similar numbers of violations which is measured by the number of demerits they have. This allows a dispatcher to prioritize high risk buffets to send out inspectors to these high-risk buffets over the lower risk buffets which can be inspected at a later time.

2. **Hierarchical Clustering – Euclidean Distance Method:**

A screenshot of a cell phone

Description automatically generatedI am going to use all 5 variables: sum\_demerits, max\_demerits, avg\_demerits, min\_demerits, and total\_inspections as the slight differences between them may be the difference that prioritizes one buffet over another as high-risk and thus would provide a better public health service. I would use 15 clusters for clustering using the Euclidean method as for each cluster the R-square increases and the variance is being better represented as the number of clusters increases. The R-square for the 15th cluster is R2=0.954 which is good.

A close up of a clock

Description automatically generated

3. **Hierarchical Clustering – City Block Method:**

For the City Block method for clustering, I also used 15 clusters to try and get a large R-square to minimize the variance. For the City Block method R2=0.963 which is marginally better than the Euclidean method. However, after the 12th cluster the change in R-square is no longer significant.

A screenshot of a cell phone

Description automatically generatedA screenshot of a cell phone

Description automatically generated

4. **K-means clustering**

Here is the code I used to standardize the data set:

proc stdize data=buffet

method=range

out=std\_lv\_inspects

outstat=stdize\_lve;

var sum\_demerits max\_demerits avg\_demerits min\_demerits total\_inspections;

run;

I would choose a K of 5 for each of the 5 variables. However, cluster 2 has only 2 clusters in it making it not very useful as a predictive value.

A screenshot of a cell phone

Description automatically generated

5. Plot your clustered results:

From the scatter plot, most of the data seems to be clustered in a positive linear way.

A close up of text on a white background

Description automatically generated

When overlaid with a map of Las Vegas itself, this geographic cluster will show where the high-risk buffets are by color coding the map, with red dots being the highest risk and green dots being the lowest risk.

A screenshot of a computer

Description automatically generated