ISE 5103 Intelligent Data Analytics

Homework 8 - Clustering

Daniel Carpenter

December 2022

Contents

1 Data

For general data preparation, please see conceptual steps below. See `.rmd` file for detailed code.

1.1 General Description

- Data used within model is from the `datasets` R package, called `ChickWeights`. Source of data located here.
- This data represents an experiment performed on 50 recently hatched chicks.
- The experimenter fed the chicks 4 separate diets while tracking their respective weights over the period of the trials.
- The four groupings of chicks had differing outcomes of weights, which can be seen later visuals.
- This model attempts to cluster the chicks based on their weight and the time performed, thus *predicting the diet fed to each.*

1.2 Data Understanding

Create a data quality report of `numeric` and `factor` data

1.2.1 Numeric Data Quality Report

Num_Numeric_Variables	Total_Observations
2	578

variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100
weight	0	1	122	71.1	35	63	103	164	373
Time	0	1	11	6.8	0	4	10	16	21

1.2.2 Factor Data Quality Report

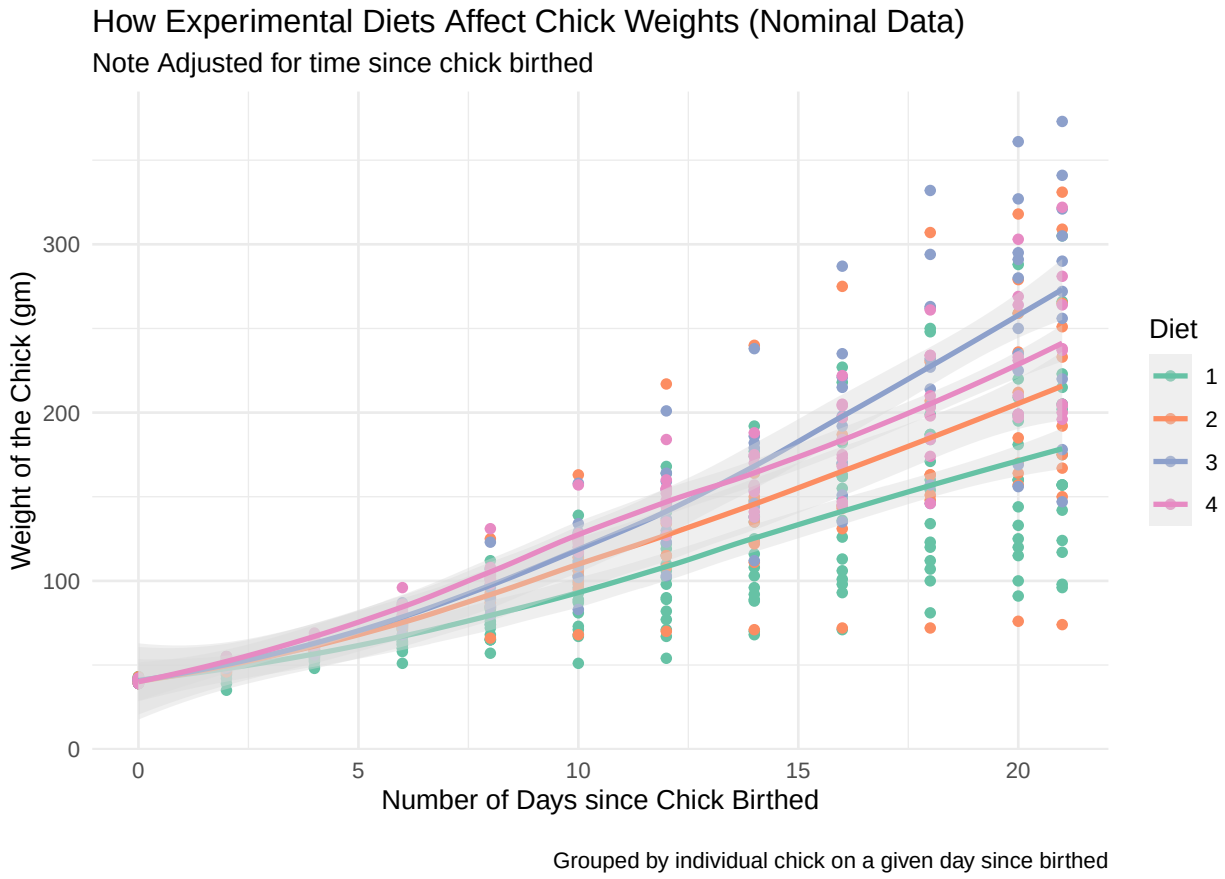
- Note that there are four distinct values within the factor field "Diet".
- Later we will attempt to replicate these 4 groupings through clustering.

Num_Factor_Variables	Total_Observations
2	578

variable	n_missing	complete_rate	n_unique	top_counts
Chick	0	1	50	13: 12, 9: 12, 20: 12, 10: 12
Diet	0	1	4	1: 220, 2: 120, 3: 120, 4: 118

1.3 Review Actual Groupings within Unadjusted, or Nominal Data

- Below shows data grouped by each chick over multiple periods of time within the study.
- The color associates with the diet few to the four groupings of chicks.
- The four lines indicate the general trend of weight gain from the diet provided to the chicks. For example,
  - Diet 1 provides the least amount of weight gain over all periods, relative to the other diet groups.
  - Diet 2 offers the second least weight gain over all periods.
  - Diet 3 and 4 stimulate similar weight gain until ~14 days since the chick hatched; however, diet 3 surpasses diet 4 after day 14.

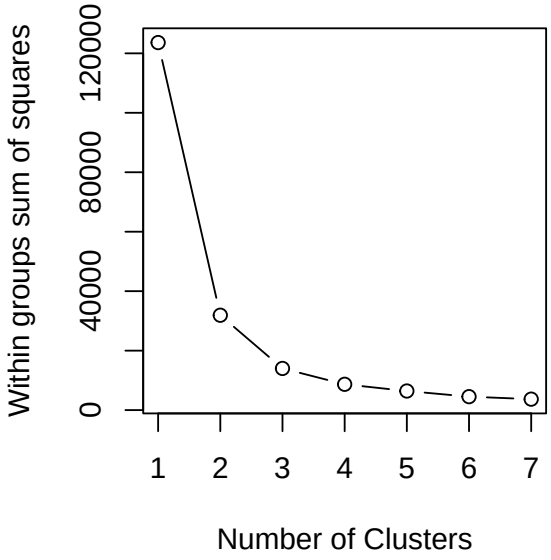


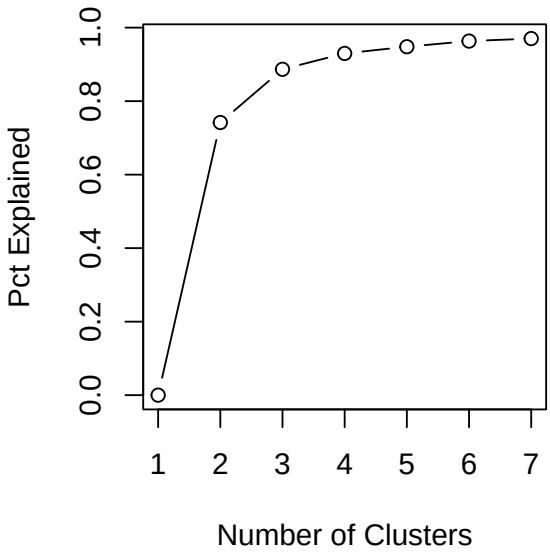
2 Perform Clustering Analysis

- Simply, the model will attempt cluster the chicks by the diet fed to them, without knowing what diet was actually given to them.
- The three models will display how the cluster of chicks' weights trend over time, as the past graph has shown.

2.1 Discover Automically Suggested Number of Clusters

- Using the "elbow" method, a plot can visually indicate the number of potential clusters that exist within the data set (assuming we do no know the actual number).
- The below elbow plots, as well the hidden results of the `NbClust` function suggest that there are around 3-5 clusters present within the data (based on the time, weight, and chick identifier).
- Knowing that there are 4 distinct clusters, as well the suggestion of the elbow point below, the k-mean and k-medoid models will attempt to discover four clusters.





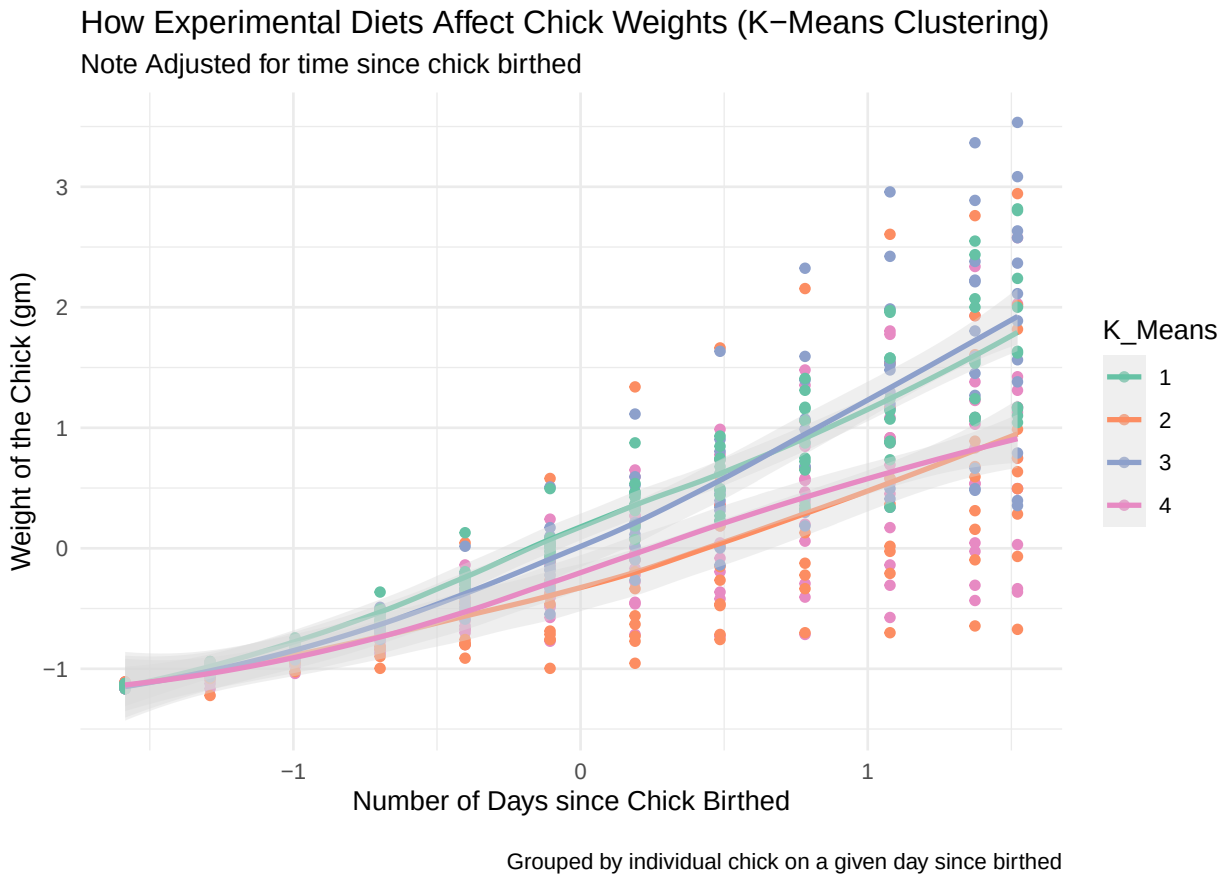
2.2 K-Means Clustering

2.2.1 *Percentage* Confusion Matrix

- Note that this confusion matrix shows the *percentage* of diet classified as the correct or incorrect class.

```
##       predicted
## actual    1    2    3    4
##      1 0.00 0.24 0.00 0.76
##      2 0.00 0.70 0.30 0.00
##      3 0.10 0.00 0.90 0.00
##      4 1.00 0.00 0.00 0.00
```

2.2.2 Visualization of Clusters



2.2.3 Interpretation

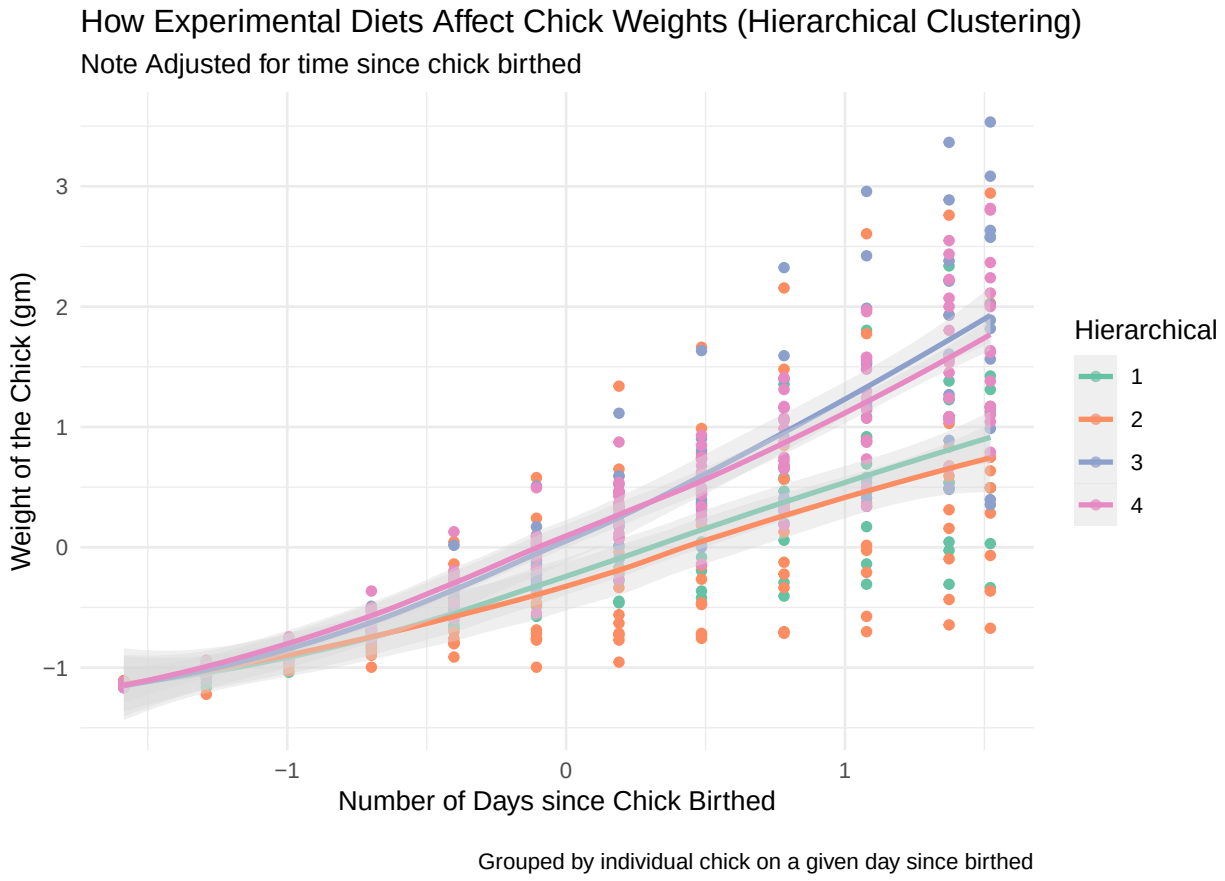
- Note that this model performs the clustering 100 times with 100 different initial seeds.
- As seen visually or within the confusion matrix, diet 1 and 4 incorrectly clustered 100% of the time. Diets 2 and 3 separated well, given by the percentage-based confusion matrix.

2.3 Hierarchical Clustering

2.3.1 *Percentage* Confusion Matrix

```
##       predicted
## actual    1    2    3    4
##      1 0.54 0.46 0.00 0.00
##      2 0.00 0.38 0.62 0.00
##      3 0.00 0.00 0.50 0.50
##      4 0.00 0.00 0.00 1.00
```

2.3.2 Visualization of Clusters



2.3.3 Interpretation

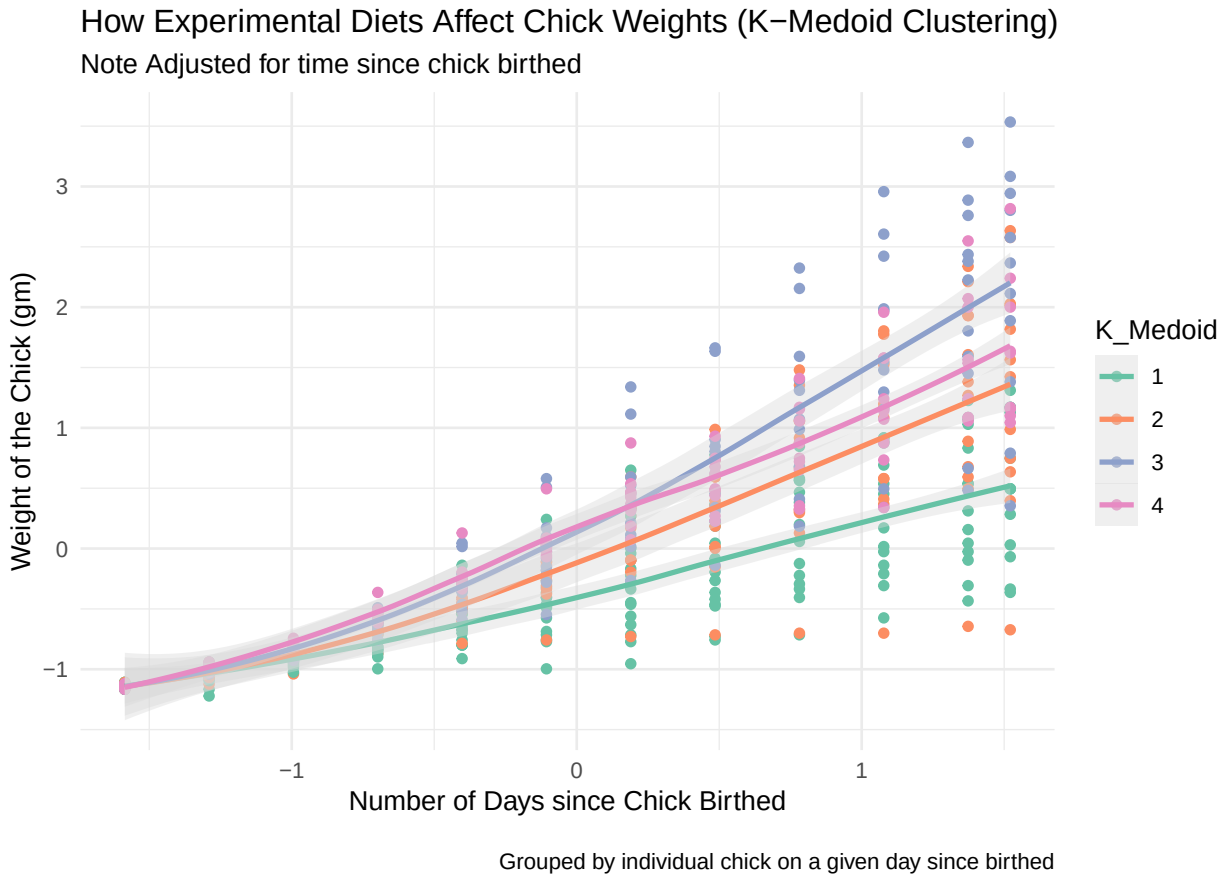
- As seen visually or within the confusion matrix, diet 1, 2, and 3 have poor levels of clustering prediction.
- Diet 4 was correctly classified 100% of the time. E.g., the model was certain when predicting that the chicks within the diet 4 group were fed diet 4.

2.4 K-Medoid Clustering

2.4.1 *Percentage* Confusion Matrix

```
##       predicted
## actual    1    2    3    4
##      1 0.84 0.16 0.00 0.00
##      2 0.00 0.90 0.10 0.00
##      3 0.00 0.00 1.00 0.00
##      4 0.00 0.00 0.00 1.00
```

2.4.2 Visualization of Clusters



2.4.3 Interpretation

- As seen visually or within the confusion matrix, the model predicted accuracy is the following:
  - Diet : 84% correct. Misclassification between diets 1 and 2.
  - Diet : 90% correct. Misclassification between diets 2 and 3
  - Diet : 100% correct
  - Diet : 100% correct
- Overall, the K-Medoid model offers the highest level of accuracy when clustering the chicks into their fed diets, given their identification, weight, and time since birthed.