ISE 5103 Intelligent Data Analytics

Homework 8 - Clustering

Daniel Carpenter

December 2022

Contents

1 General Data Prep

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

1.1 Read Training Data

Clean data to ensure each read variable has the correct data type (factor, numeric, Date, etc.)

1.2 Create `numeric` and `factor` *base* `data frames`

Make data set of `numeric` variables called `df.base.numeric`

Make data set of `factor` variables called `df.base.factor`

2 Data Understanding

Create a data quality report of `numeric` and `factor` data
Created function called `dataQualityReport()` to create factor and numeric QA report

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

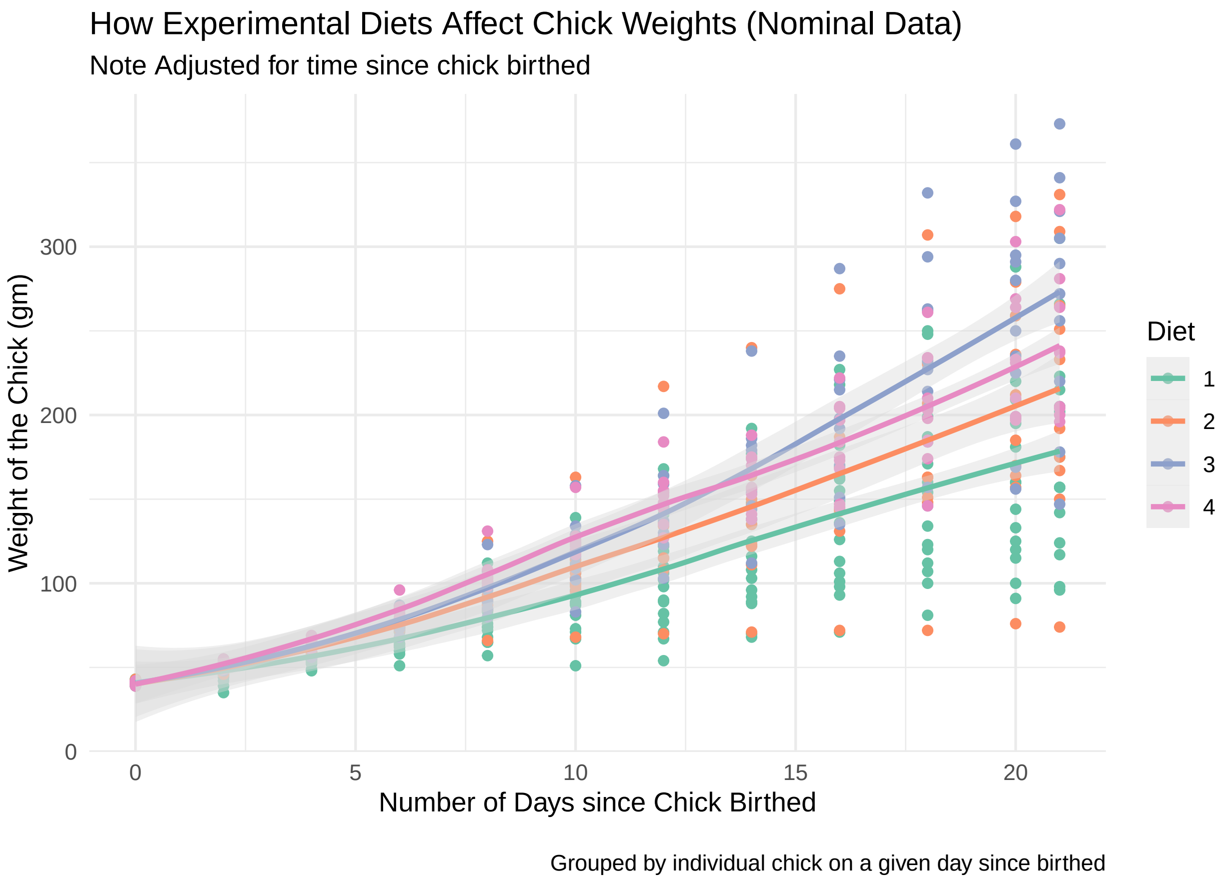
2.2 Factor Data Quality Report

- 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

2.3 Review Actual Groupings within Unadjusted, or Nominal Data



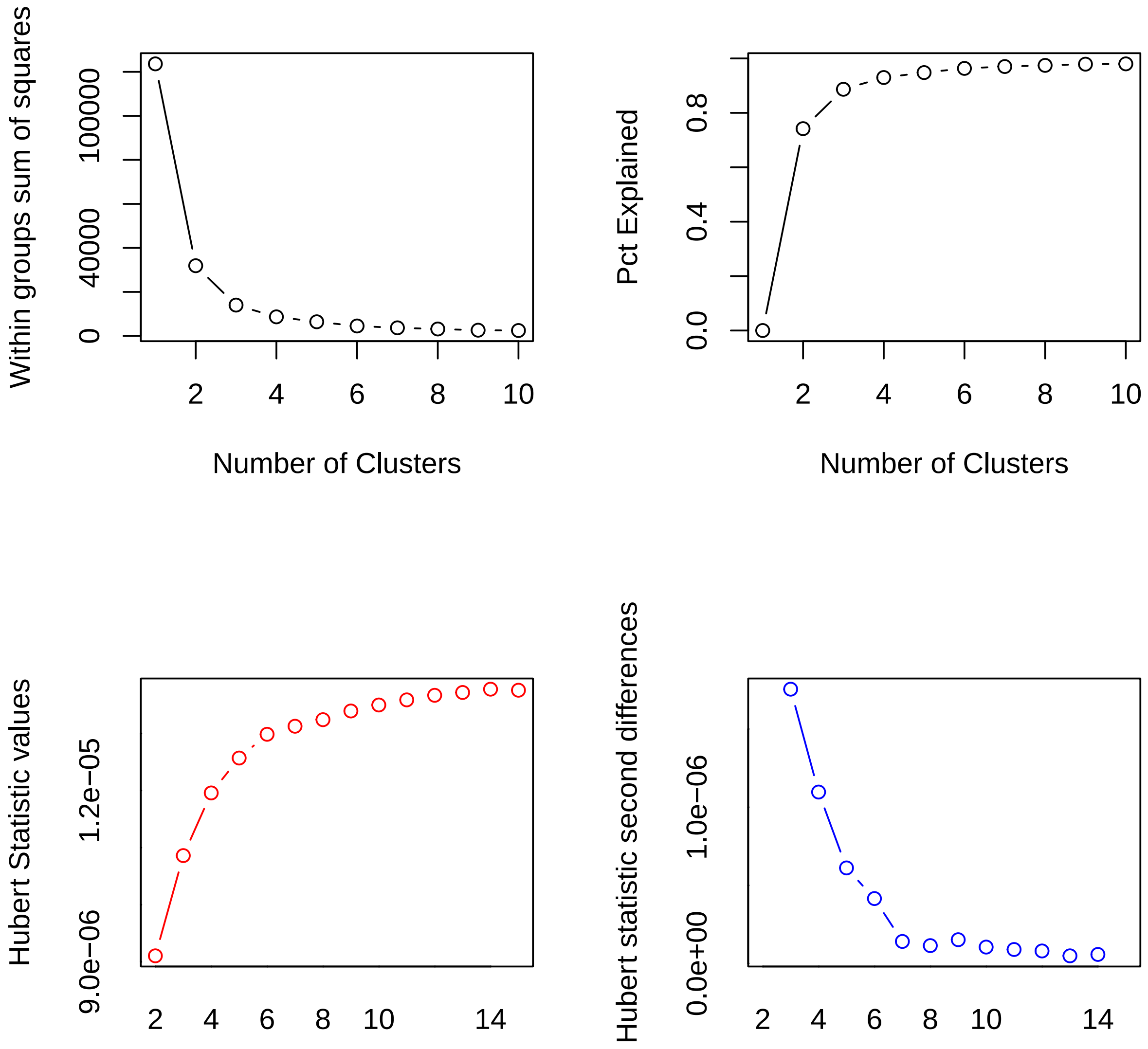
How Experimental Diets Affect Chick Weights (Nominal Data)

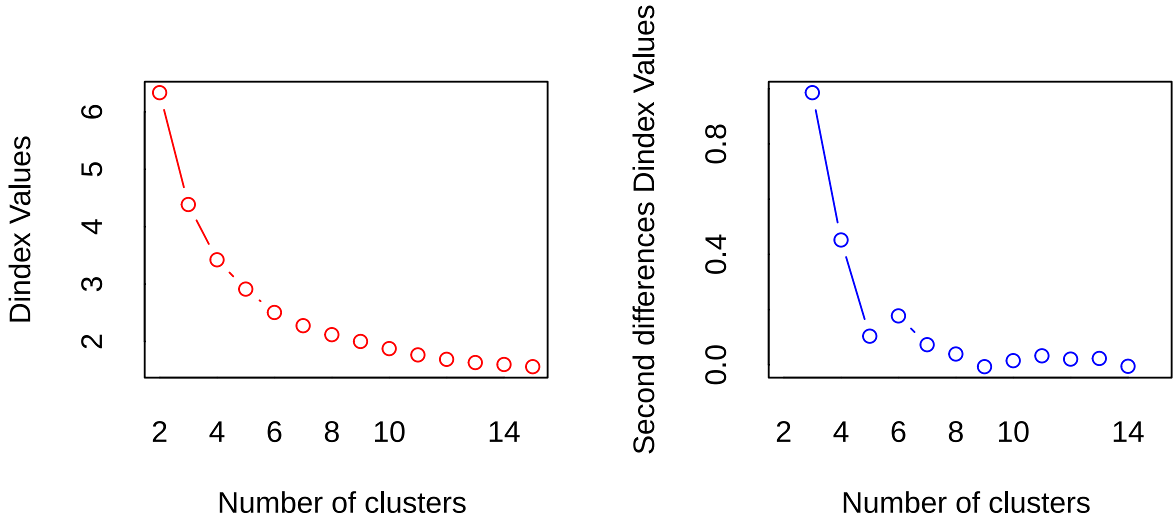
Note Adjusted for time since chick birthed

Grouped by individual chick on a given day since birthed

2.4 Clustering Analysis

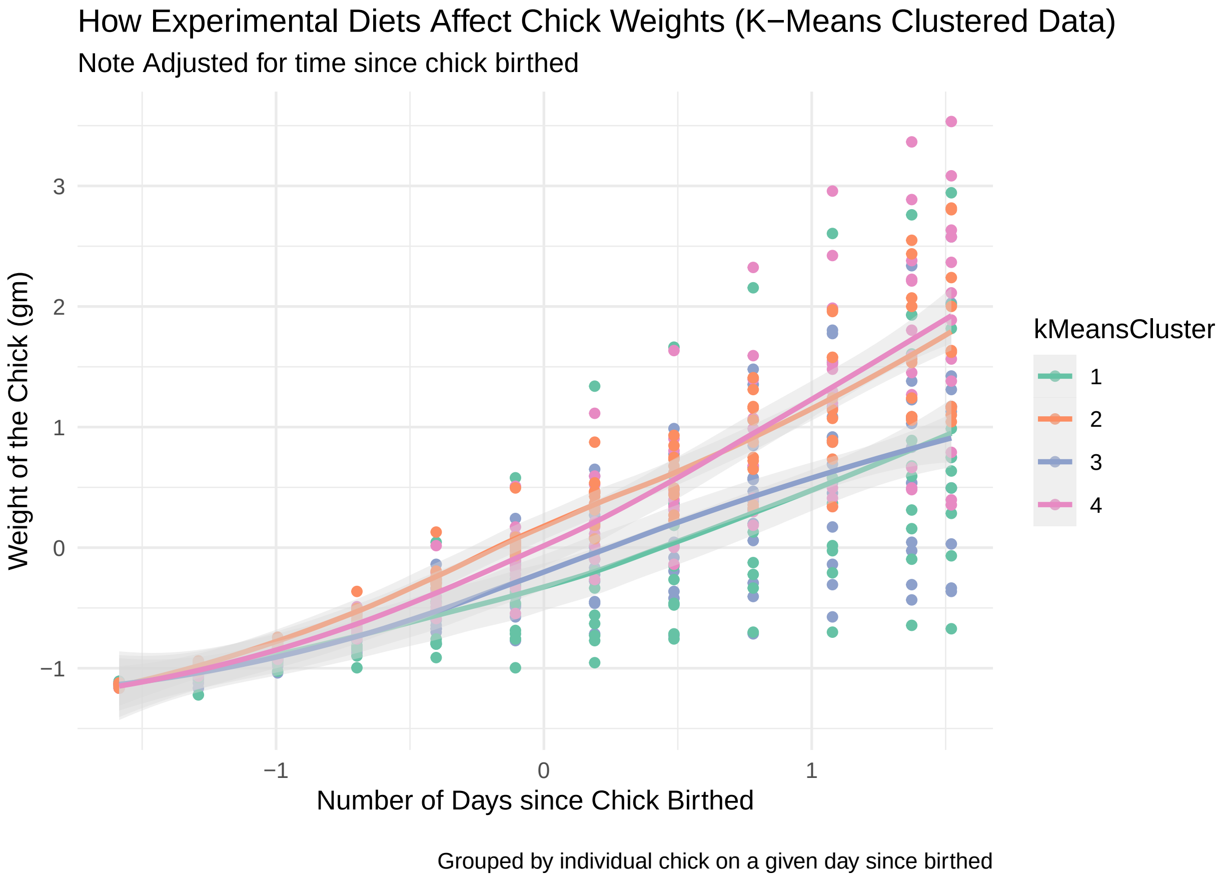
2.4.1 Discover Automically Suggested Number of Clusters




Dindex Values
2
3
4
5
6
2
4
6
8
10
14
Number of clusters
Second differences Dindex Values
0.0
0.4
0.8
2
4
6
8
10
14
Number of clusters

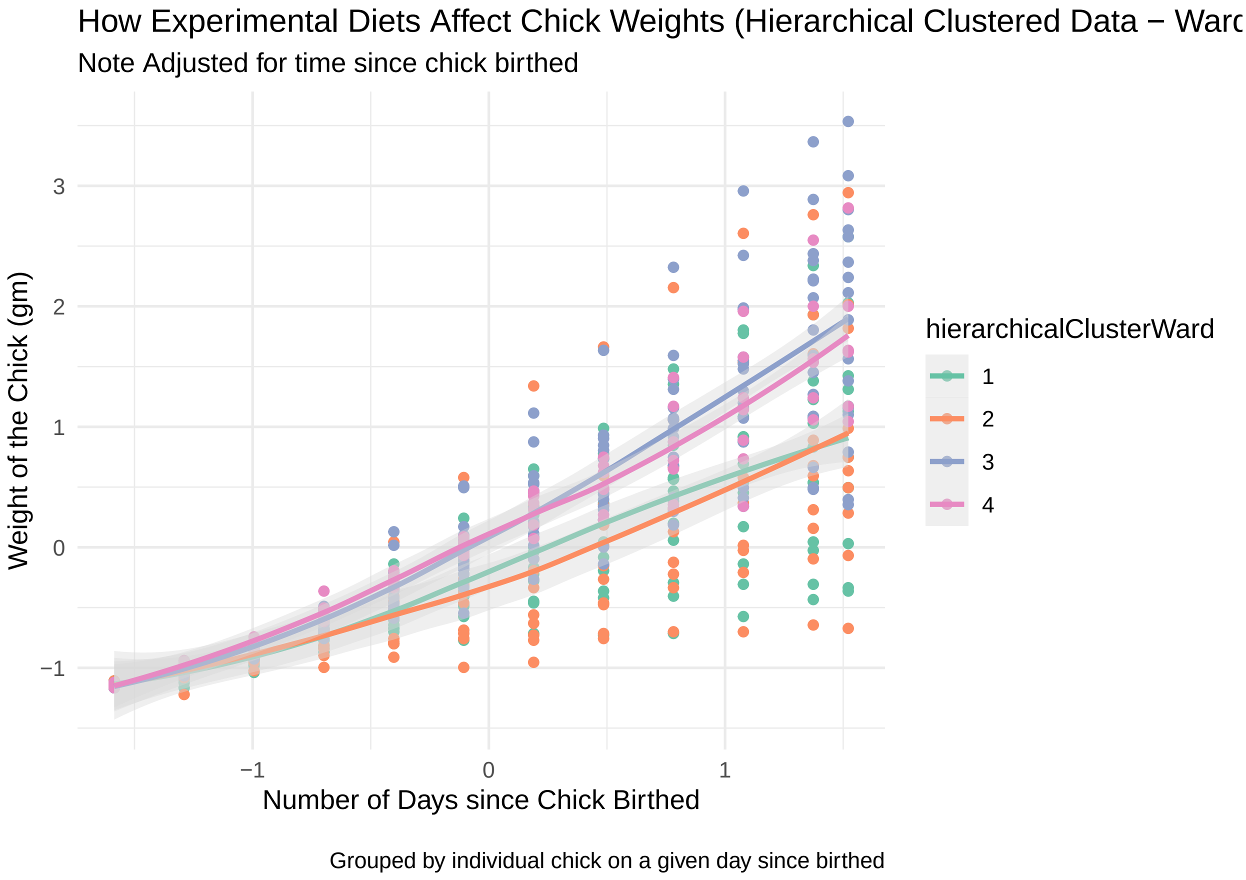
2.4.2 K-Means Clustering

```
##
##          1   2   3   4
##    1    53   0 167   0
##    2    84   0   0  36
##    3     0  12   0 108
##    4     0 118   0   0
```



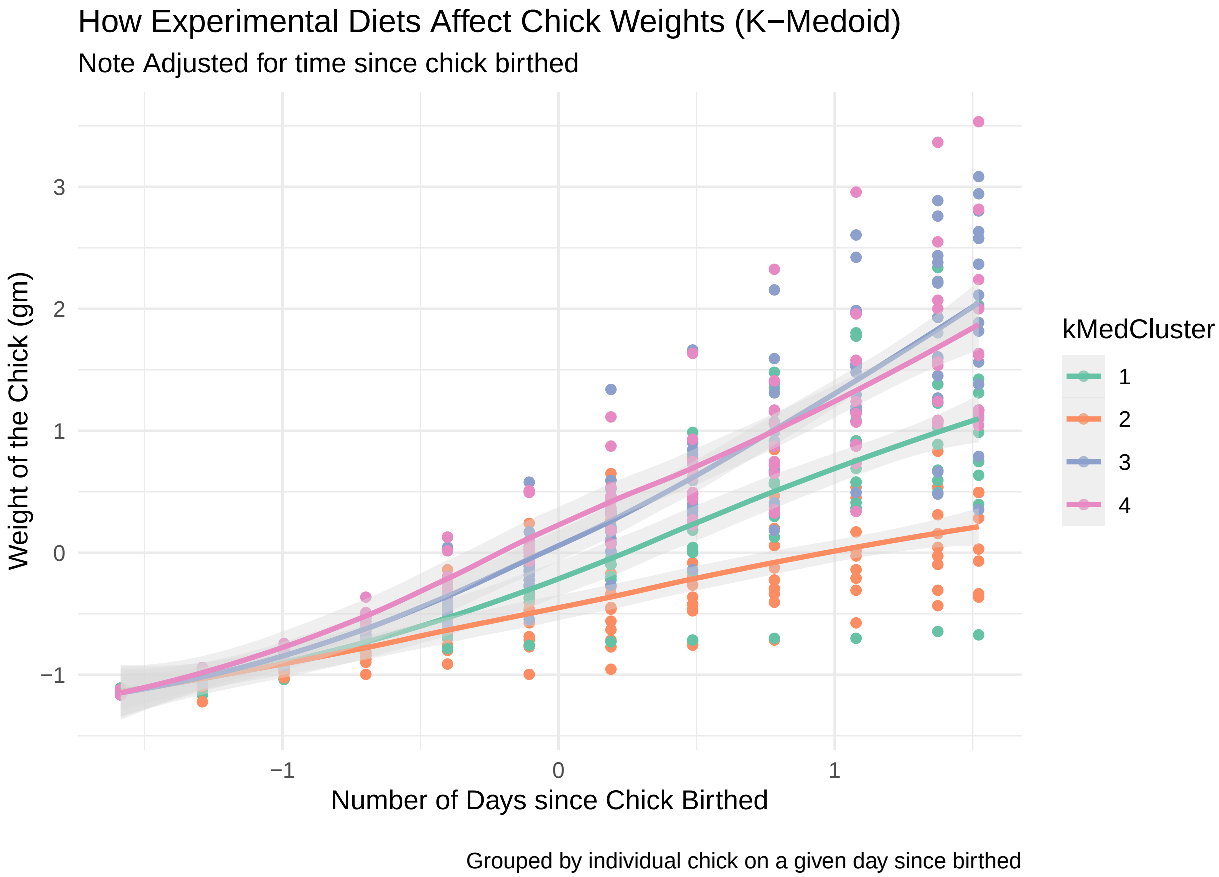
2.4.3 Hierarchical Clustering

```
##
##         1   2   3   4
##     1 167  53   0   0
##     2   0  84  36   0
##     3   0   0 120   0
##     4   0   0  32  86
```



2.4.4 K-Medoid Clustering

```
##
##       1   2   3   4
##   1  84 136   0   0
##   2  60   0  60   0
##   3   0   0 108  12
##   4   0   0   0 118
```


How Experimental Diets Affect Chick Weights (K−Medoid)
Note Adjusted for time since chick birthed
Weight of the Chick (gm)
Number of Days since Chick Birthed
kMedCluster
Grouped by individual chick on a given day since birthed