

ISE 5103 Intelligent Data Analytics

Homework 8 - Clustering

Daniel Carpenter

December 2022

Contents

1	General Data Prep	2
1.1	Read Training Data	2
1.2	Create numeric and factor <i>base data frames</i>	2
2	Data Understanding	2
2.1	Numeric Data Quality Report	2
2.2	Factor Data Quality Report	2
2.3	Review Target Cluster	3
2.4	Clustering Analysis	4

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

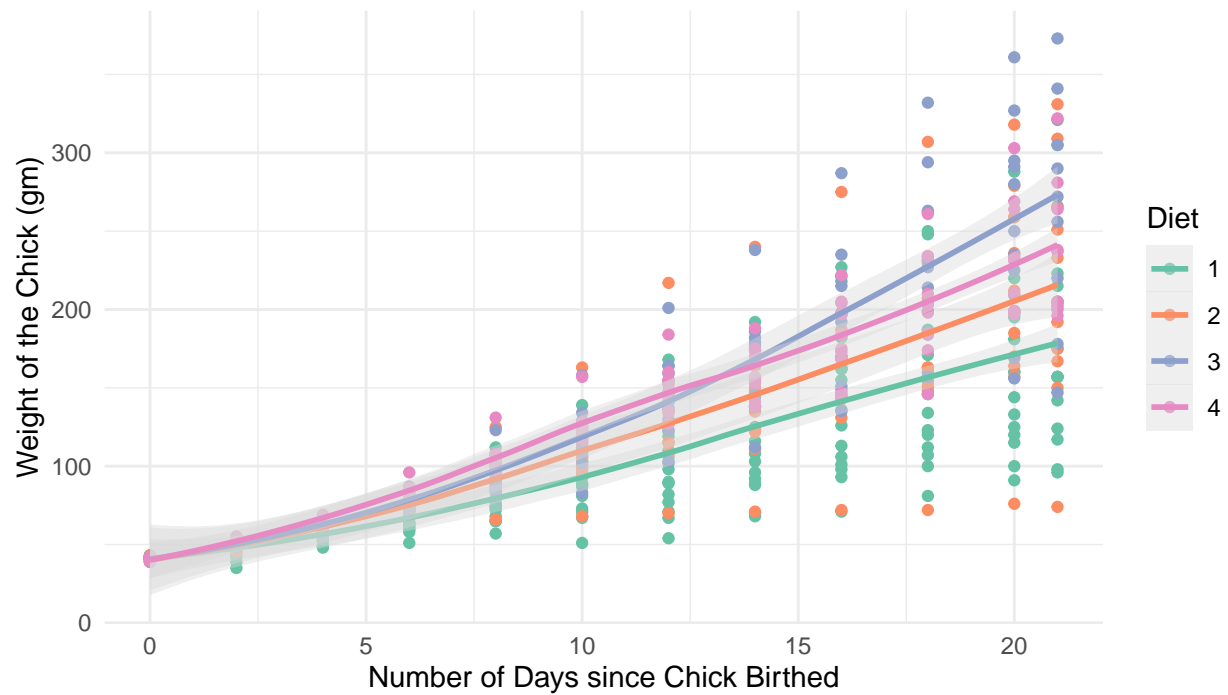
Num_Factor_Variables	Total_Observations
2	578

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 Target Cluster

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