Cracking Boston Crab

Presented by Thomas $ @ 2018-09-07

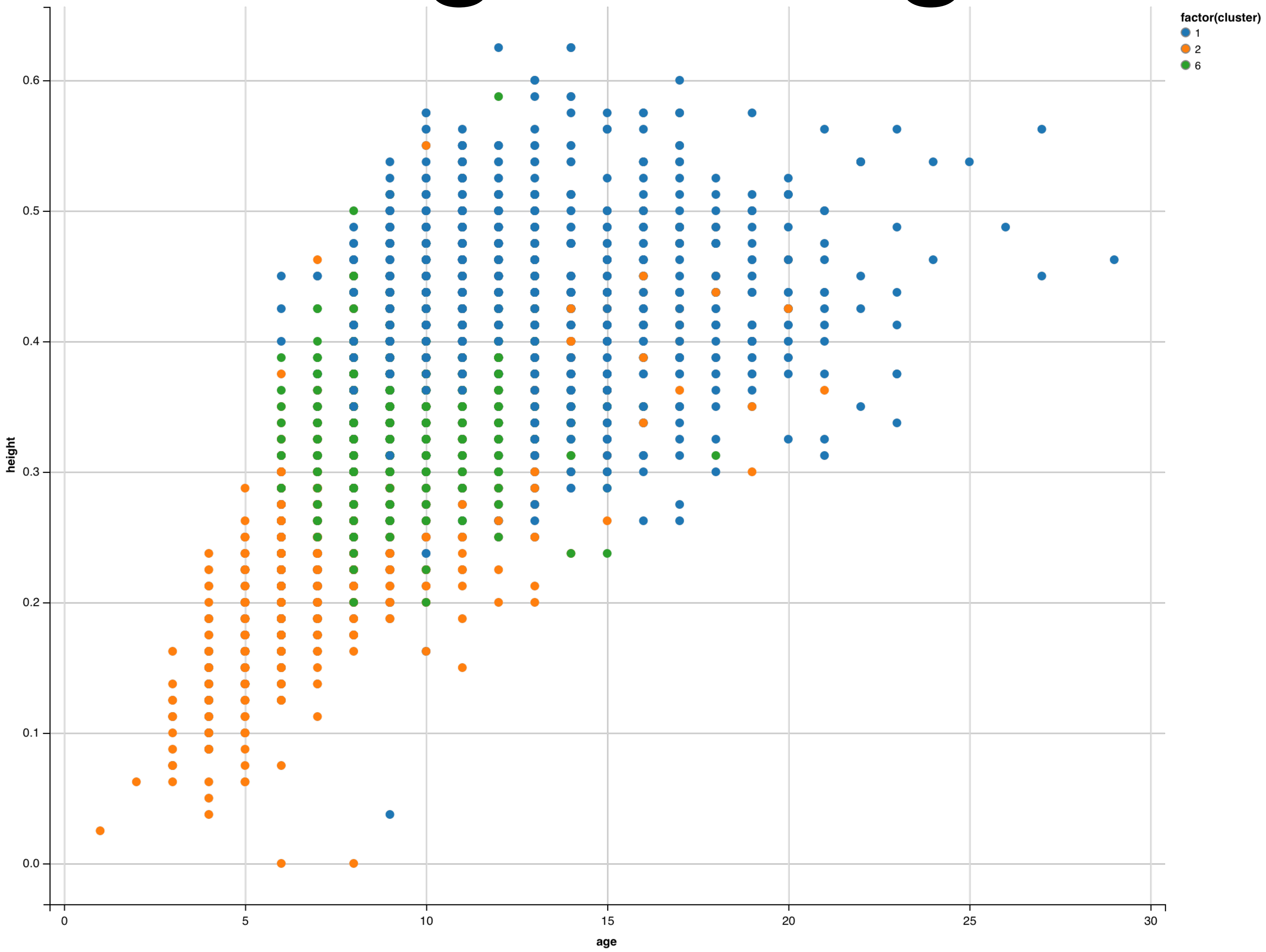
Objective

- Simply predict the age of the crab in boston based on their size, weight and diameters etc

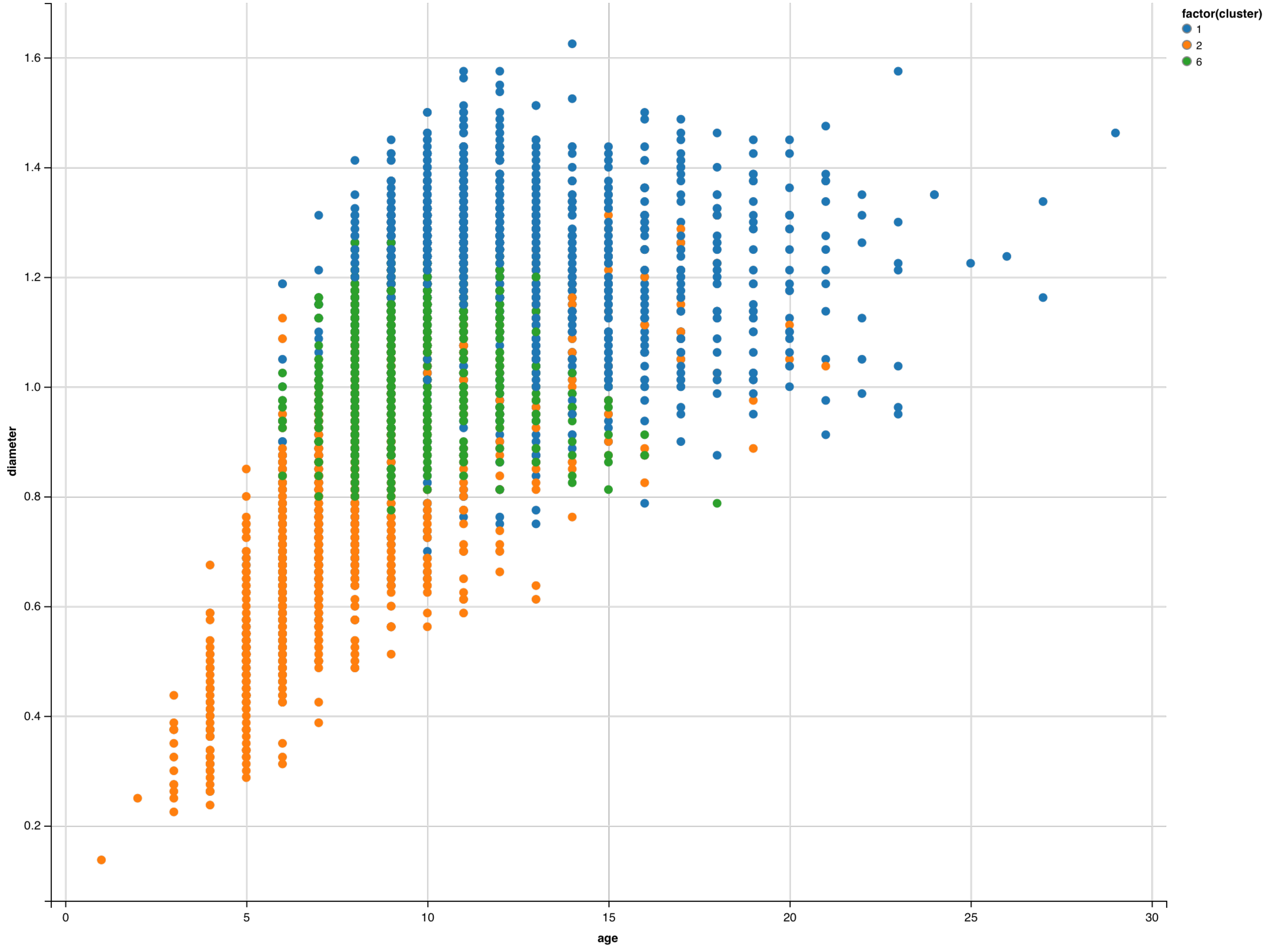
Challenges

- Read files from pdf with dynamic structure and dirty data
- Lack of strong correlation from the features to crab's age
- Without additional information, assume there are more than one spices in the data set, so further unsupervised k-mean is performed
- Data set split 70% (Train) / 30% (Final)
  within the 70% Train > 60% (Test) / 40% (Valid)

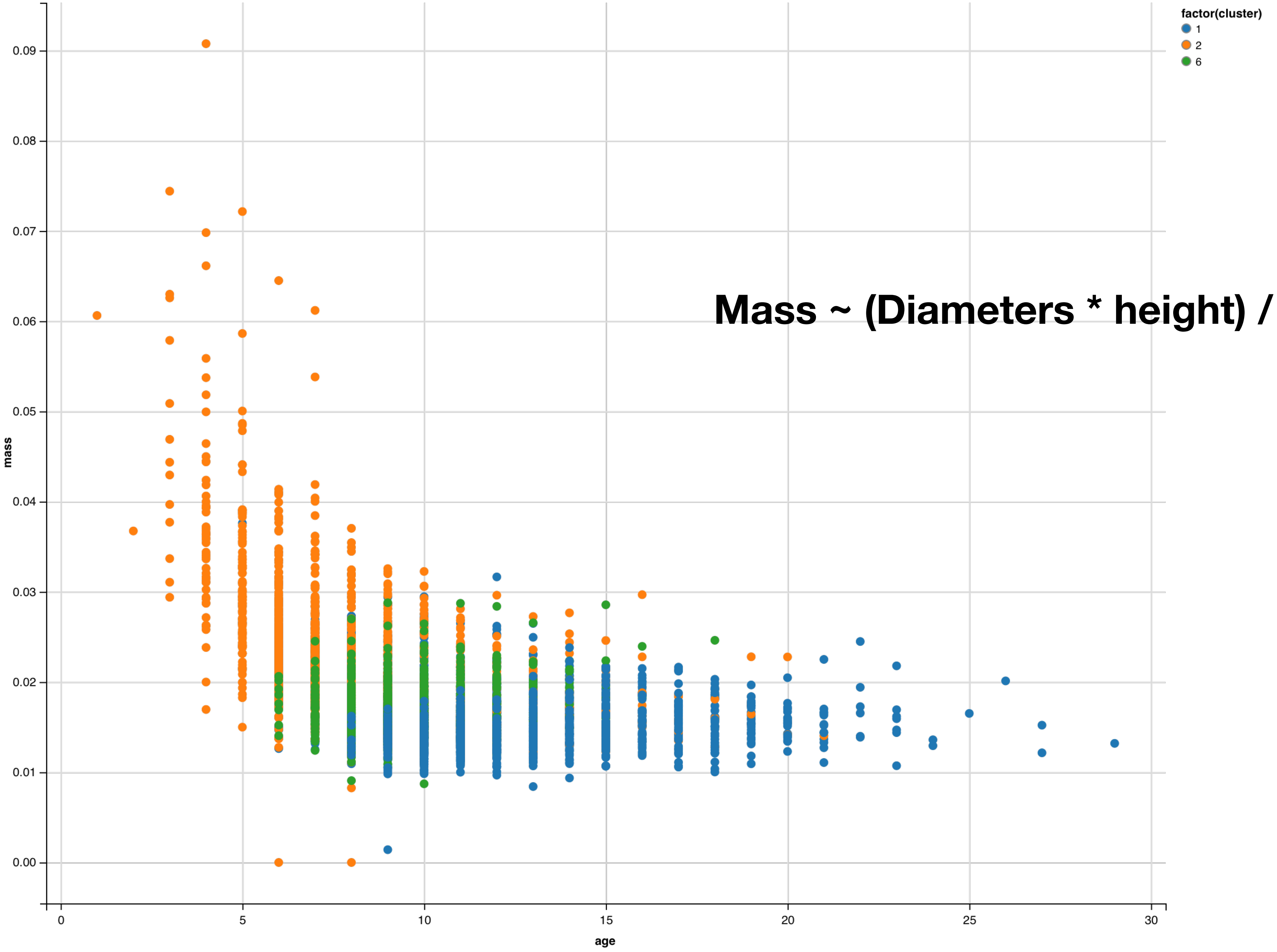
Height vs Age



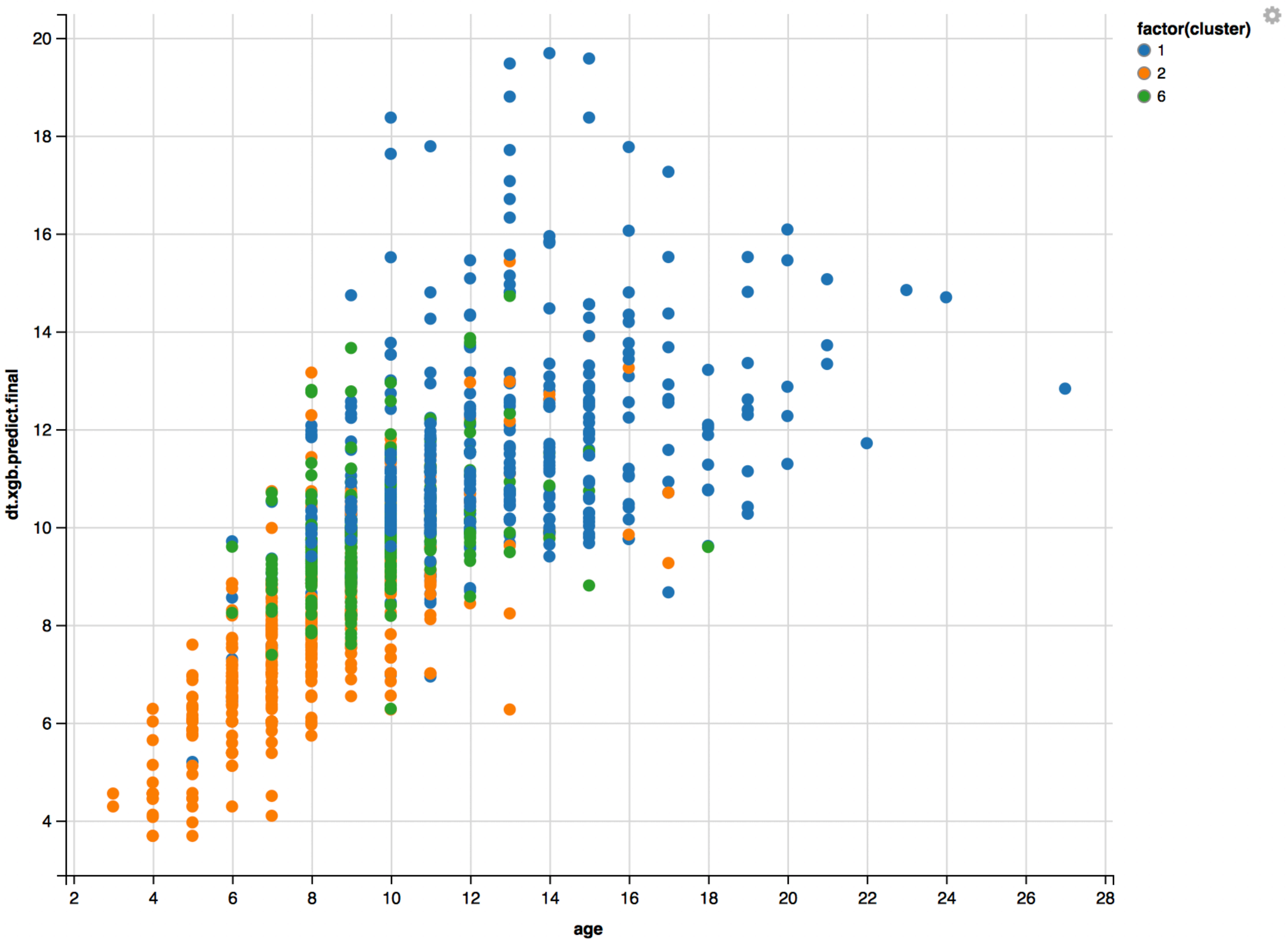
Diameters vs Age



Mass vs Age



Mass ~ (Diameters * height) / weight


factor(cluster)
● 1
● 2
● 6
dt.xgb.predict.final
4
6
8
10
12
14
16
18
20
age
2
4
6
8
10
12
14
16
18
20
22
24
26
28

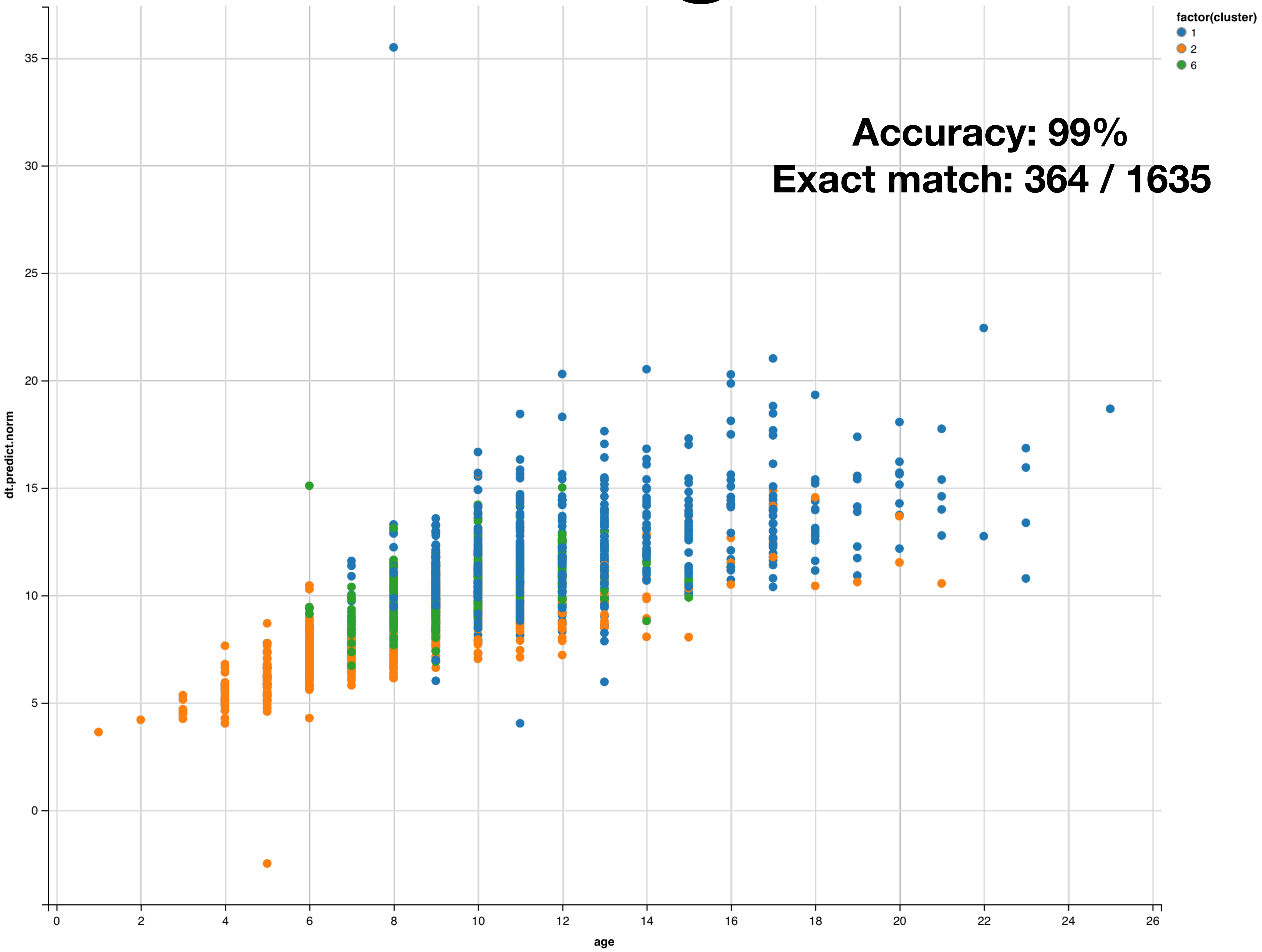
ML Modeling

- 2 types of ML models are used in this analysis
  - Linear regression models
  - Xgboost

ML Modeling

- 2 types of ML models are used in this analysis
    - Linear regression models
    - Xgboost
- Accuracy defined as:
  1 - (abs(Actual Age - Predicted Age) / Actual Age)
- Exact Match
  Sum(Round(Predicted Age) = Age) / Total # of Results

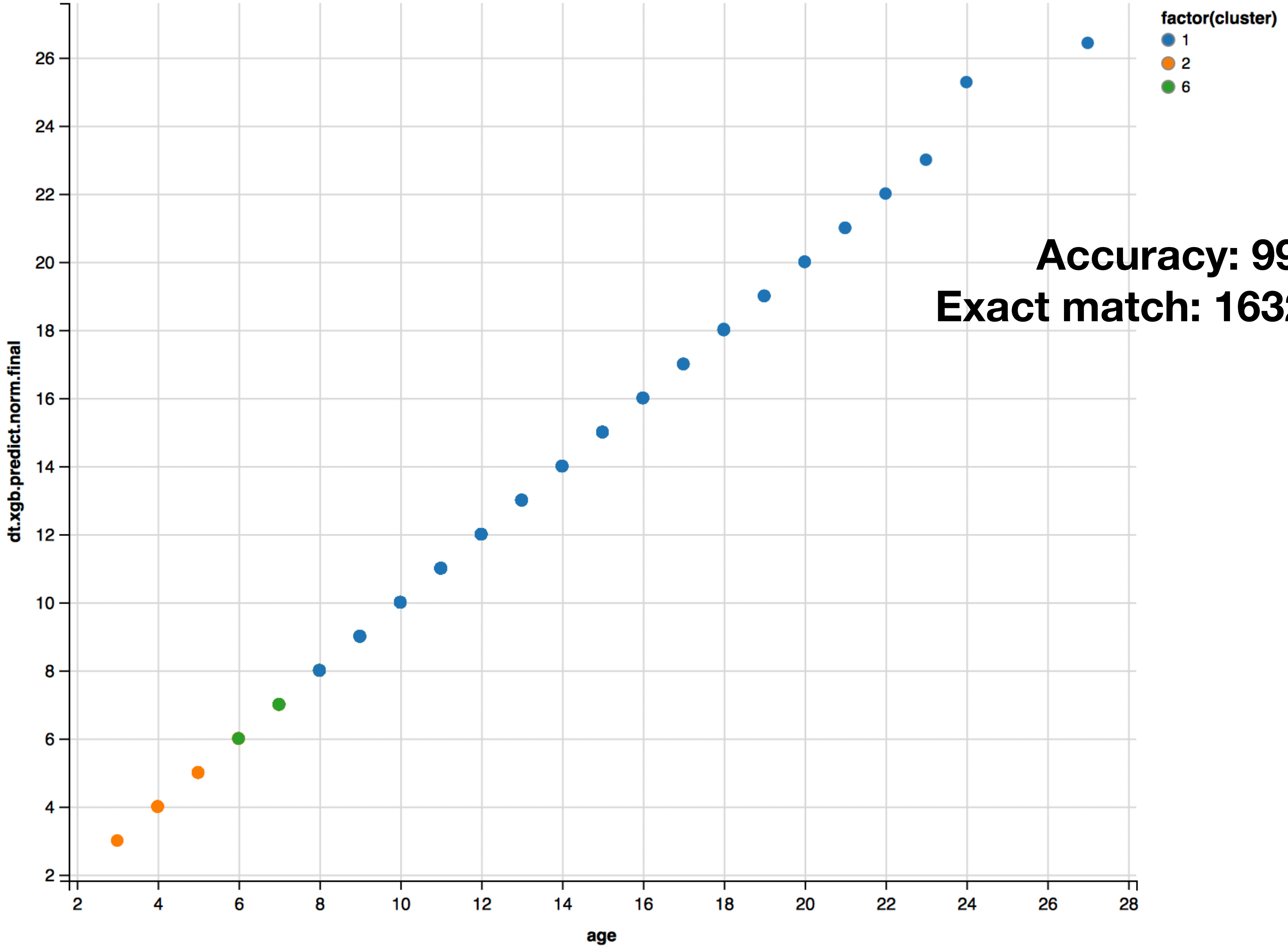
Linear Regression



Accuracy: 99%

Exact match: 364 / 1635

Xgboost



Accuracy: 99%
Exact match: 1632 / 1635

Conclusion

- In net Xgboost was able to achieve great accuracy in both valid and final data sets and scored exceptionally high in the accuracy and # of exact matched evaluation criteria.

- As such the artifacts of XGboost has been selected as the core model for the program for prediction in rScript which can be used for MI reporting, Data ETL.