CRAB AGE

GROUP PROJECT BY

CHIARA ANNI

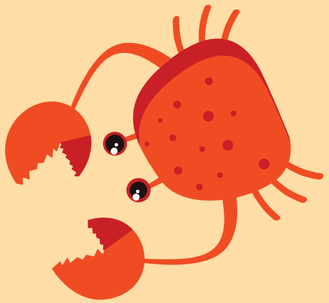
GUGLIELMO BERZANO

ALESSANDRO CANTONI

MATTEO MATONE

MICHELA MAZZAGLIA

RICCARDO STURLA



GENERAL INFO

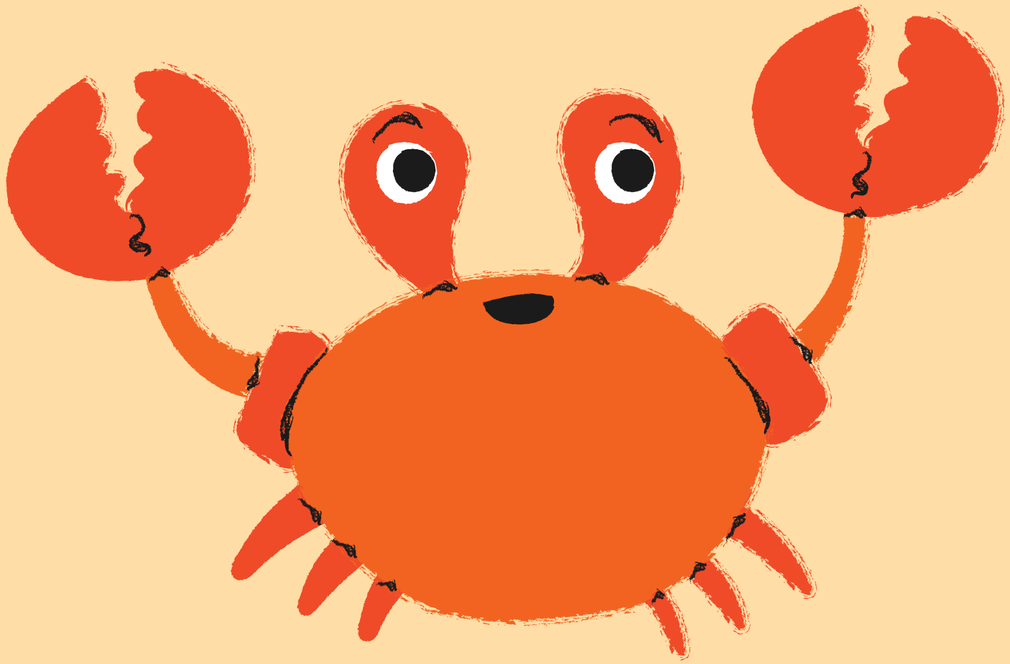
The dataset can be found here

The code is published on our Github profiles, feel free to look it up!

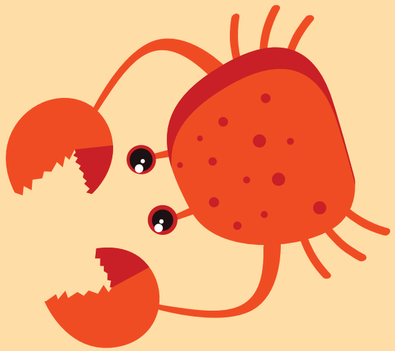
OUR GOAL

Predict life expectancy of crabs based on characteristics of their body

LOOK AT OUR DATASET!



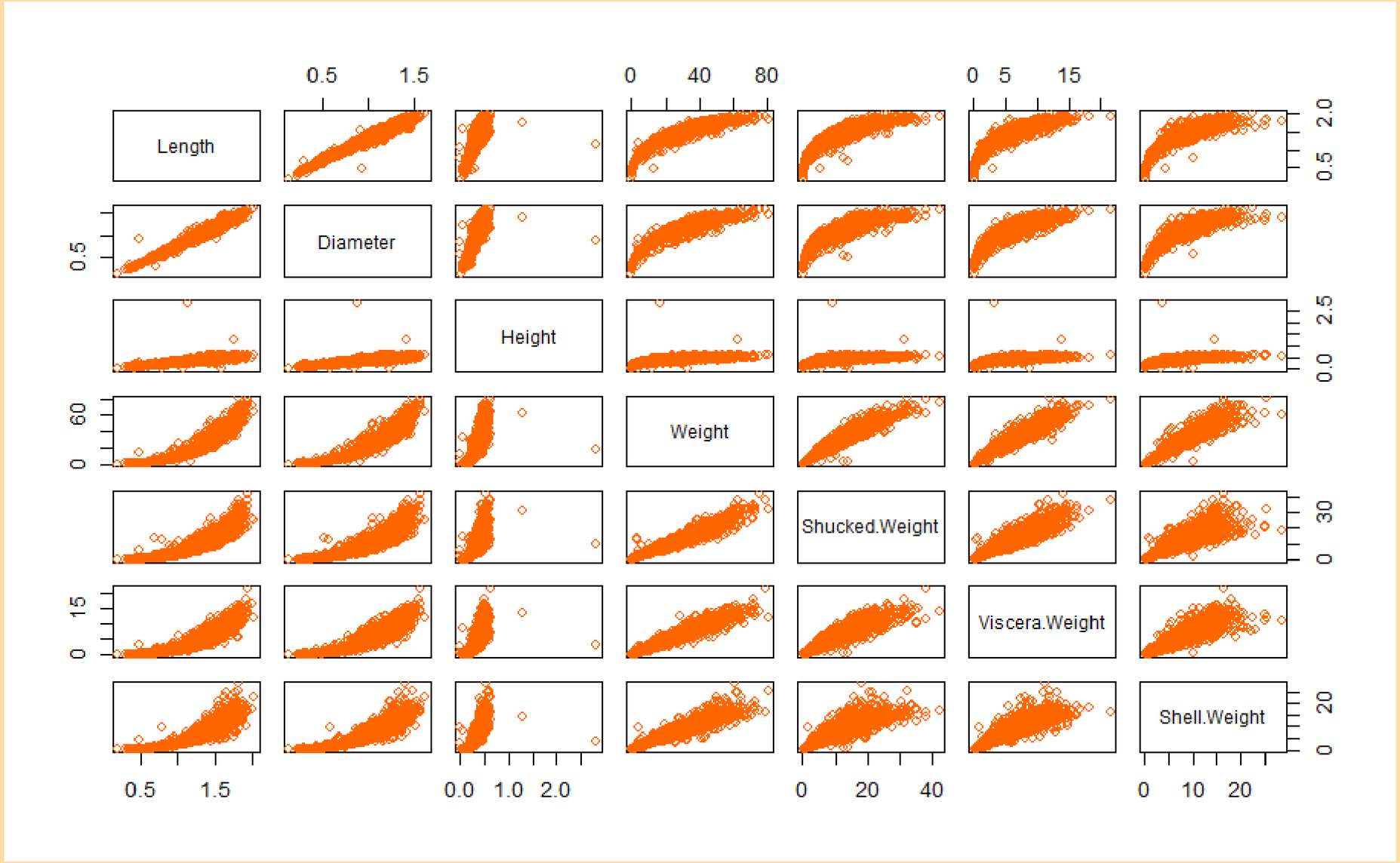
HEAD(DS)



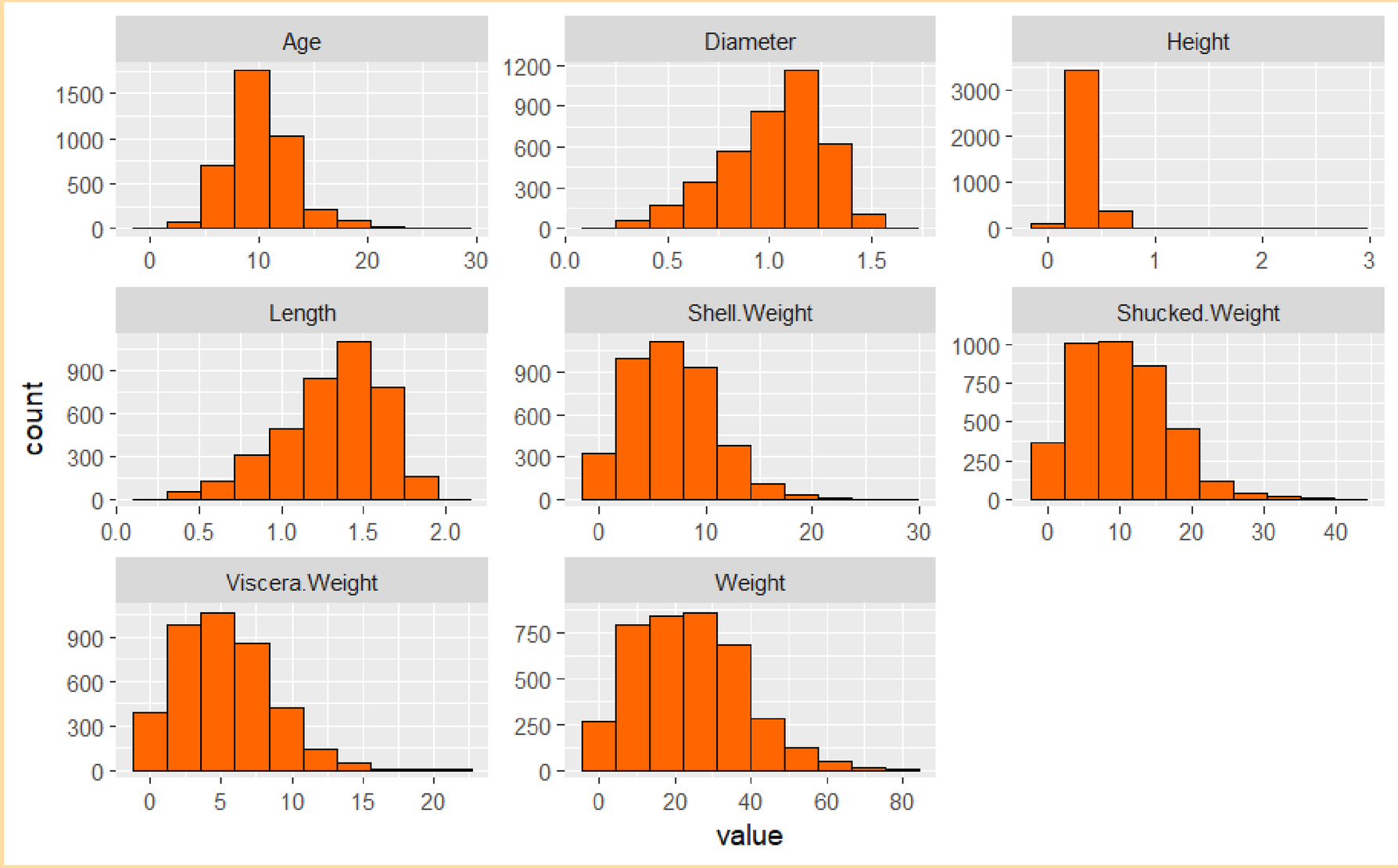
	Sex	Length	Diameter	Height	Weight	Shucked.Weight	Viscera.Weight	Shell.Weight	Age
1	F	1.4375	1.1750	0.4125	24.635715	12.332033	5.584852	6.747181	9
2	M	0.8875	0.6500	0.2125	5.400580	2.296310	1.374951	1.559222	6
3	I	1.0375	0.7750	0.2500	7.952035	3.231843	1.601747	2.764076	6
4	F	1.1750	0.8875	0.2500	13.480187	4.748541	2.282135	5.244657	10
5	I	0.8875	0.6625	0.2125	6.903103	3.458639	1.488349	1.700970	6

9 variables, 3893 observations

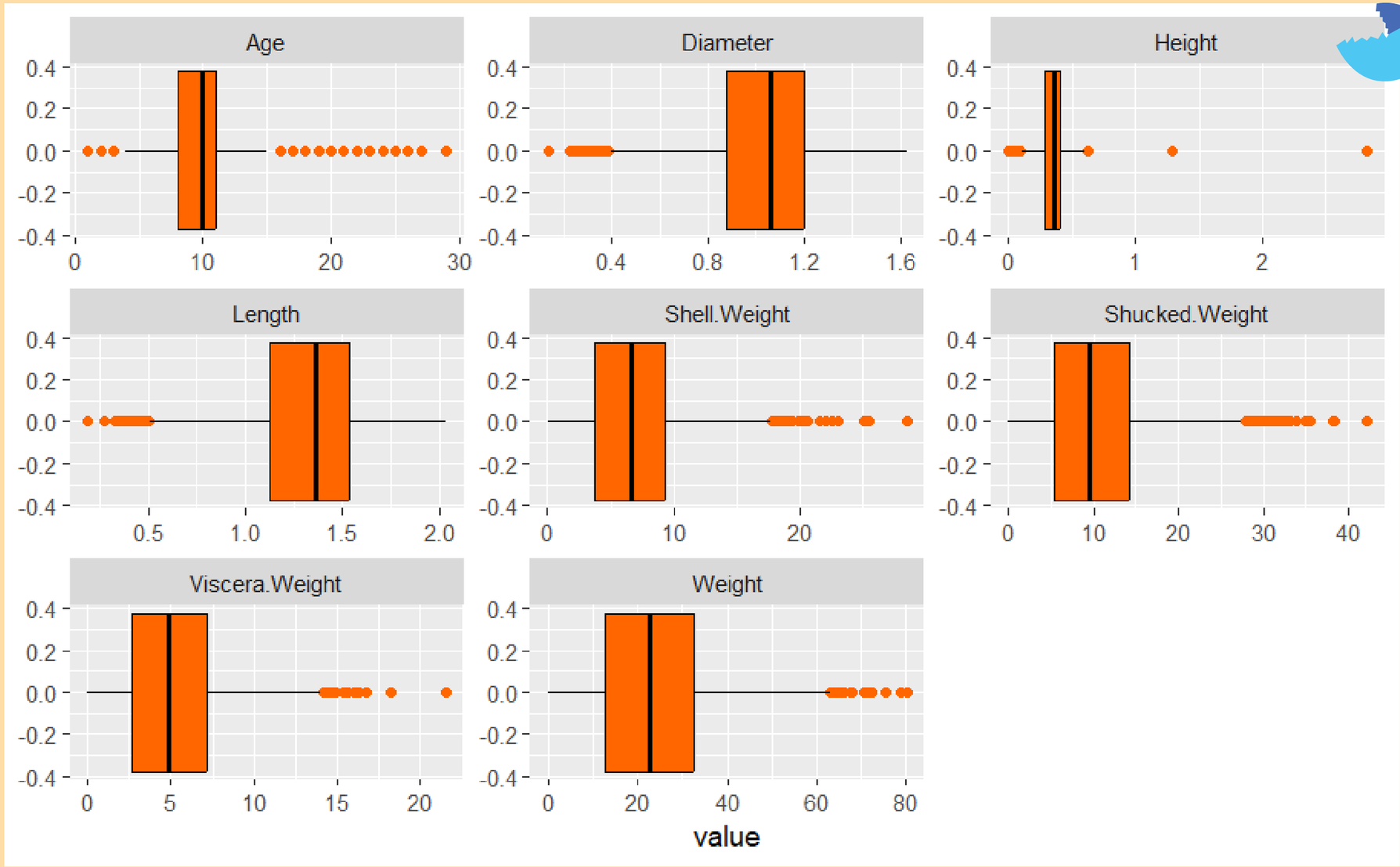
EXPLORATORY ANALYSIS, CORRELATIONS



EXPLORATORY ANALYSIS, HISTOGRAMS



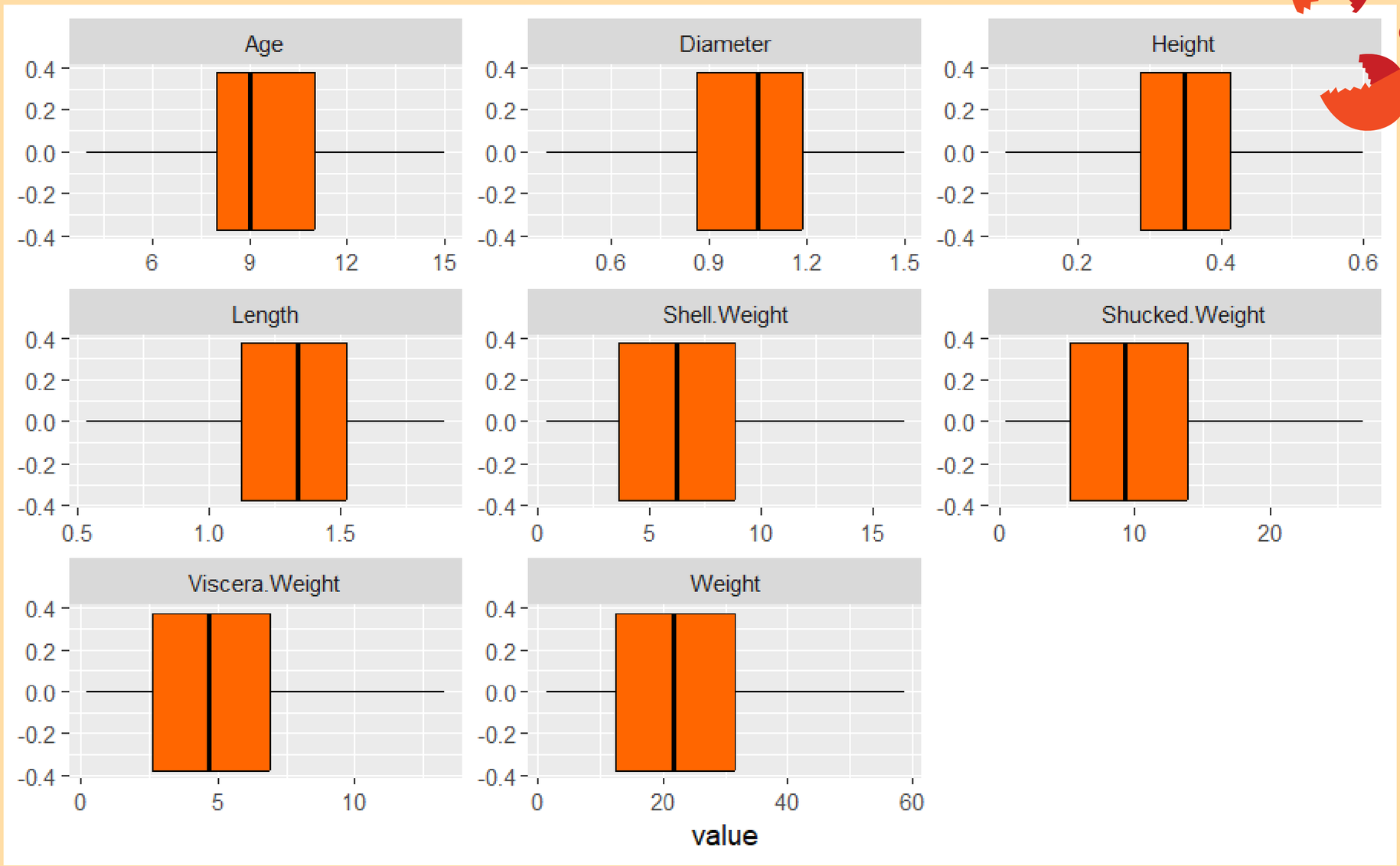
EXPLORATORY ANALYSIS, BOXPLOTS



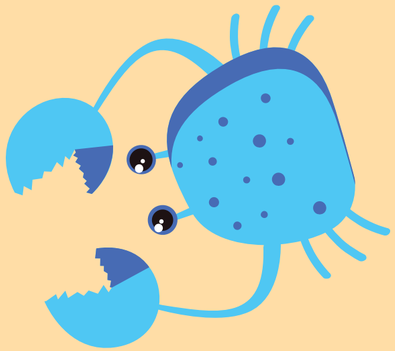
HOW TO DEAL WITH OUTLIERS?

Remove them!

OUTLIER REMOVAL



HEAD(DS)



Split the variable Sex into three dummies

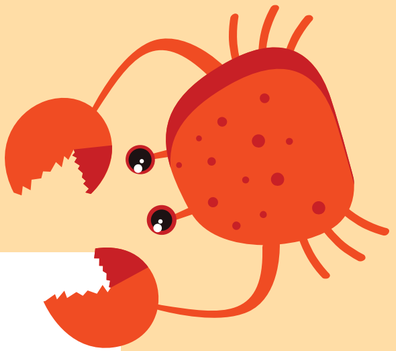
	Sex	Length	Diameter	Height	Weight	Shucked.Weight	Viscera.Weight	Shell.Weight	Age	Sex_M	Sex_F	Sex_I
1	F	1.4375	1.1750	0.4125	24.635715	12.332033	5.584852	6.747181	9	0	1	0
2	M	0.8875	0.6500	0.2125	5.400580	2.296310	1.374951	1.559222	6	1	0	0
3	I	1.0375	0.7750	0.2500	7.952035	3.231843	1.601747	2.764076	6	0	0	1
4	F	1.1750	0.8875	0.2500	13.480187	4.748541	2.282135	5.244657	10	0	1	0
5	I	0.8875	0.6625	0.2125	6.903103	3.458639	1.488349	1.700970	6	0	0	1

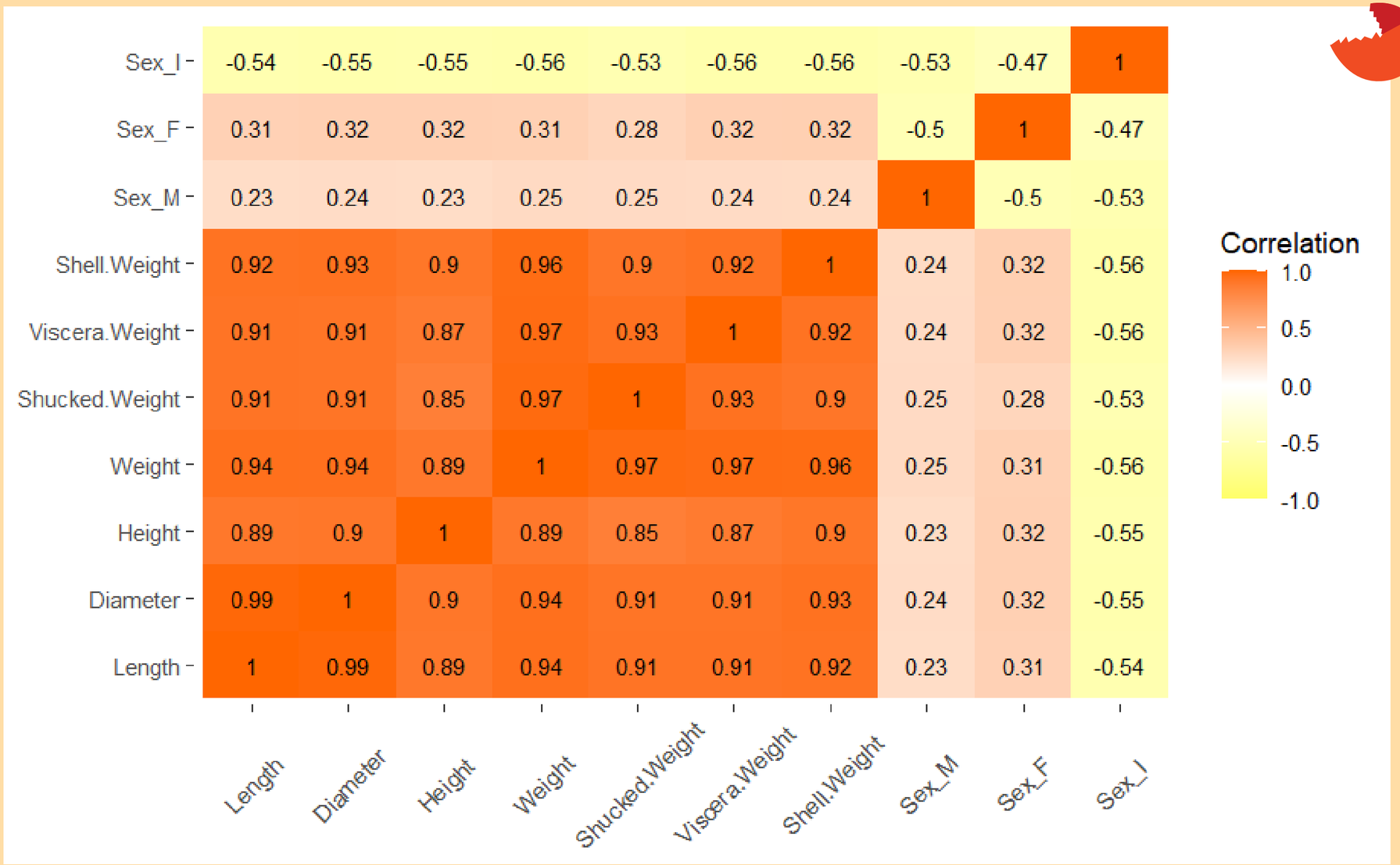
12 variables, 3509 observations

Perfect multicollinearity?

Dummy variable trap?

HEATMAP FOR CORRELATION





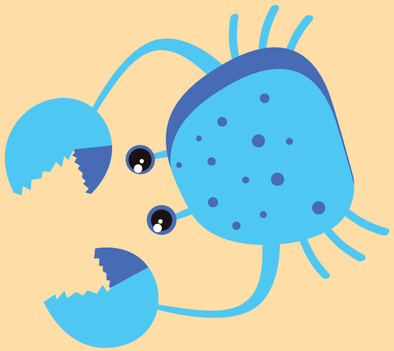
SUPERVISED LEARNING

REGRESSION

Relationships between a dependent variable and one or more independent variables.

LINEAR REGRESSION

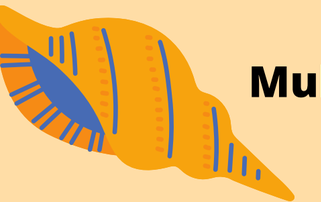
Preliminar analysis



Call: `lm(formula = Age ~ . - Sex - Sex_F, data = ds1)`

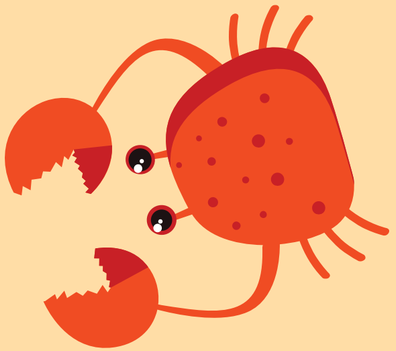
Coefficients:

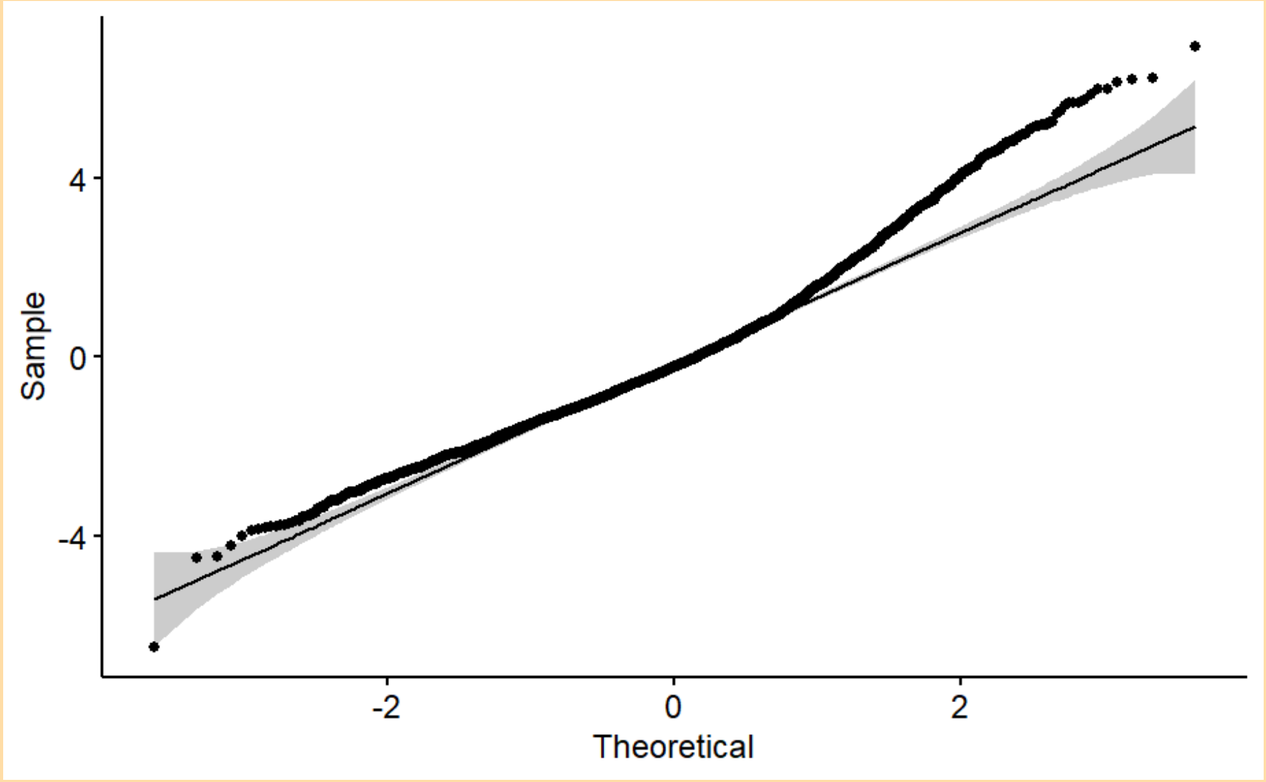
	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	3.90141	0.27948	13.959	< 2e-16	***
Length	-0.06865	0.62768	-0.109	0.912919	
Diameter	4.03443	0.76713	5.259	1.53e-07	***
Height	6.08191	0.80516	7.554	5.37e-14	***
Weight	0.14934	0.02306	6.477	1.06e-10	***
Shucked.Weight	-0.44680	0.02639	-16.931	< 2e-16	***
Viscera.Weight	-0.14179	0.04050	-3.501	0.000469	***
Shell.Weight	0.23075	0.03846	6.000	2.18e-09	***
Sex_M	0.07990	0.06856	1.165	0.243951	
Sex_I	-0.81966	0.07279	-11.261	< 2e-16	***

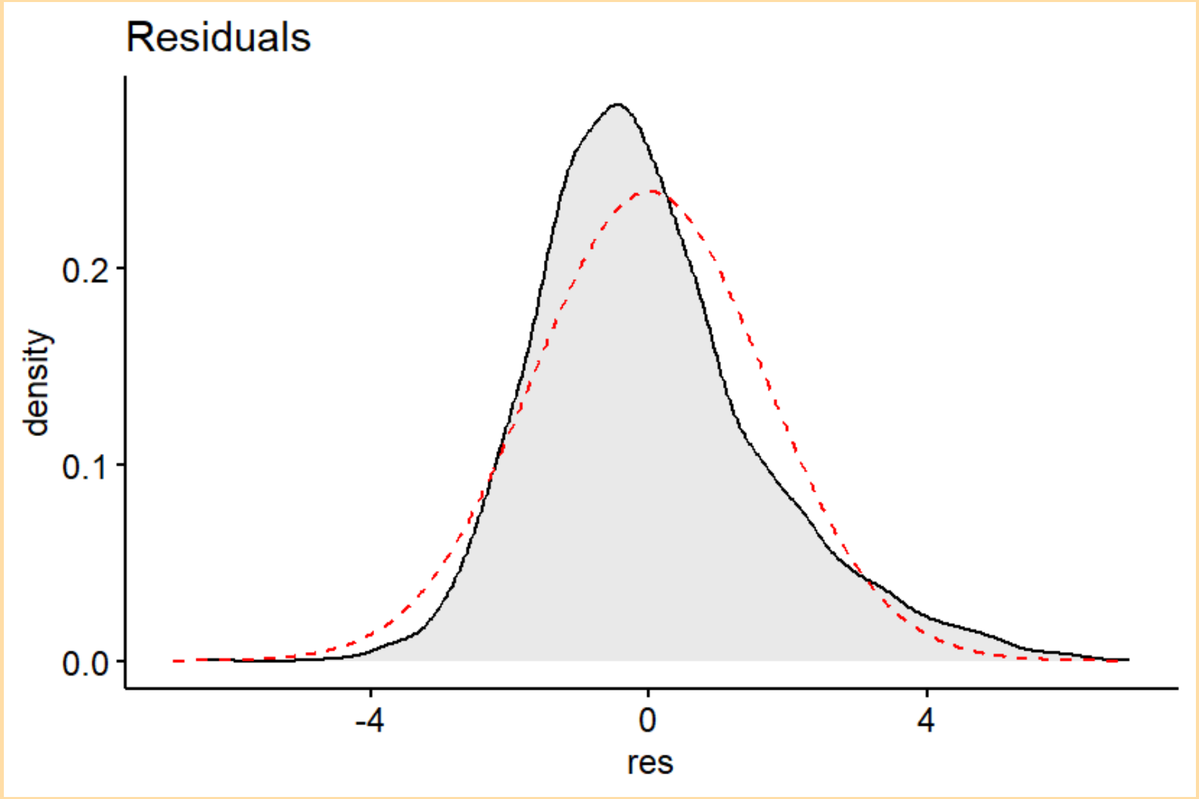


Multiple R-squared: 0.5029, Adjusted R-squared: 0.5018

CHECK THE RESIDUALS TO DETECT OUTLYING Y OBSERVATIONS



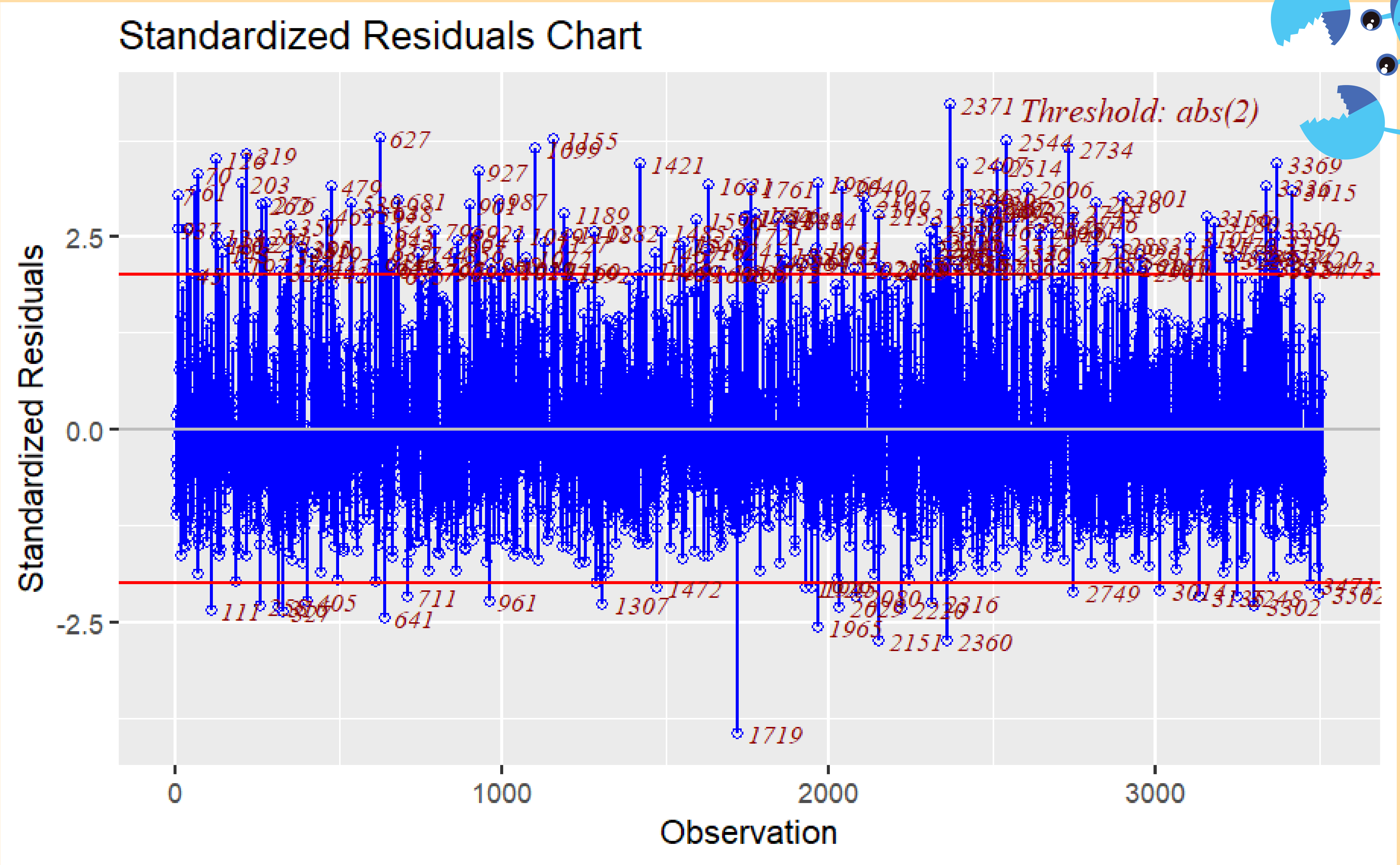




Distribution and Density

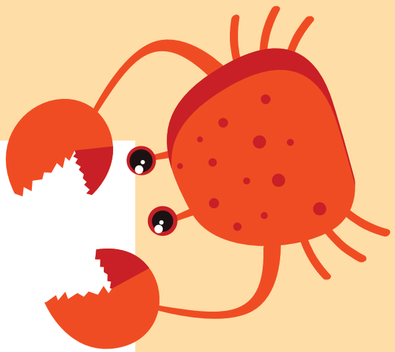
SHAPIRO-WILK NORMALITY TEST

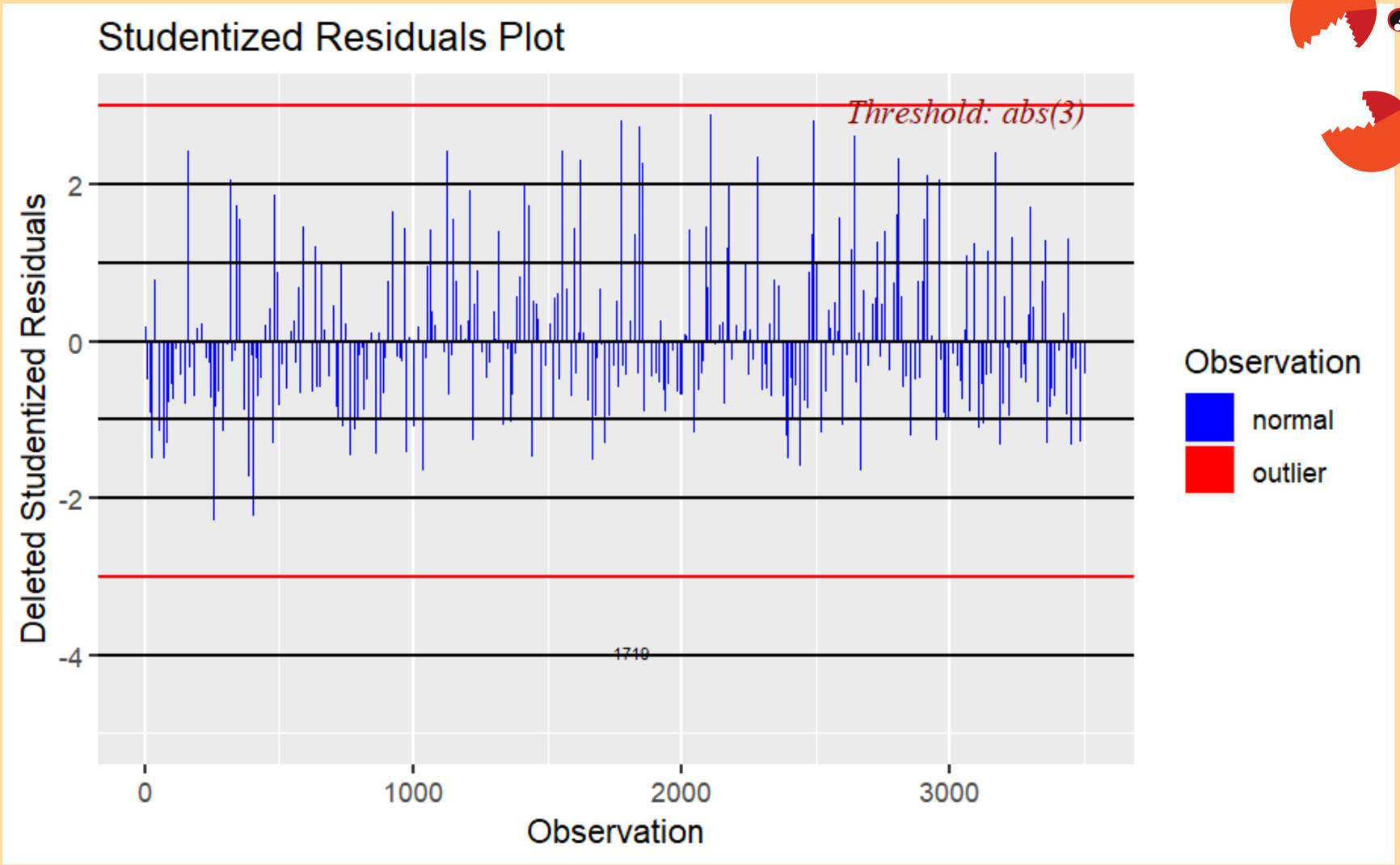
W = 0.92731, P-VALUE < 2.2E-16



Standardized Residuals Chart
Threshold: abs(2)
Standardized Residuals
Observation
2.5
0.0
-2.5
0
1000
2000
3000
2371
627
1719

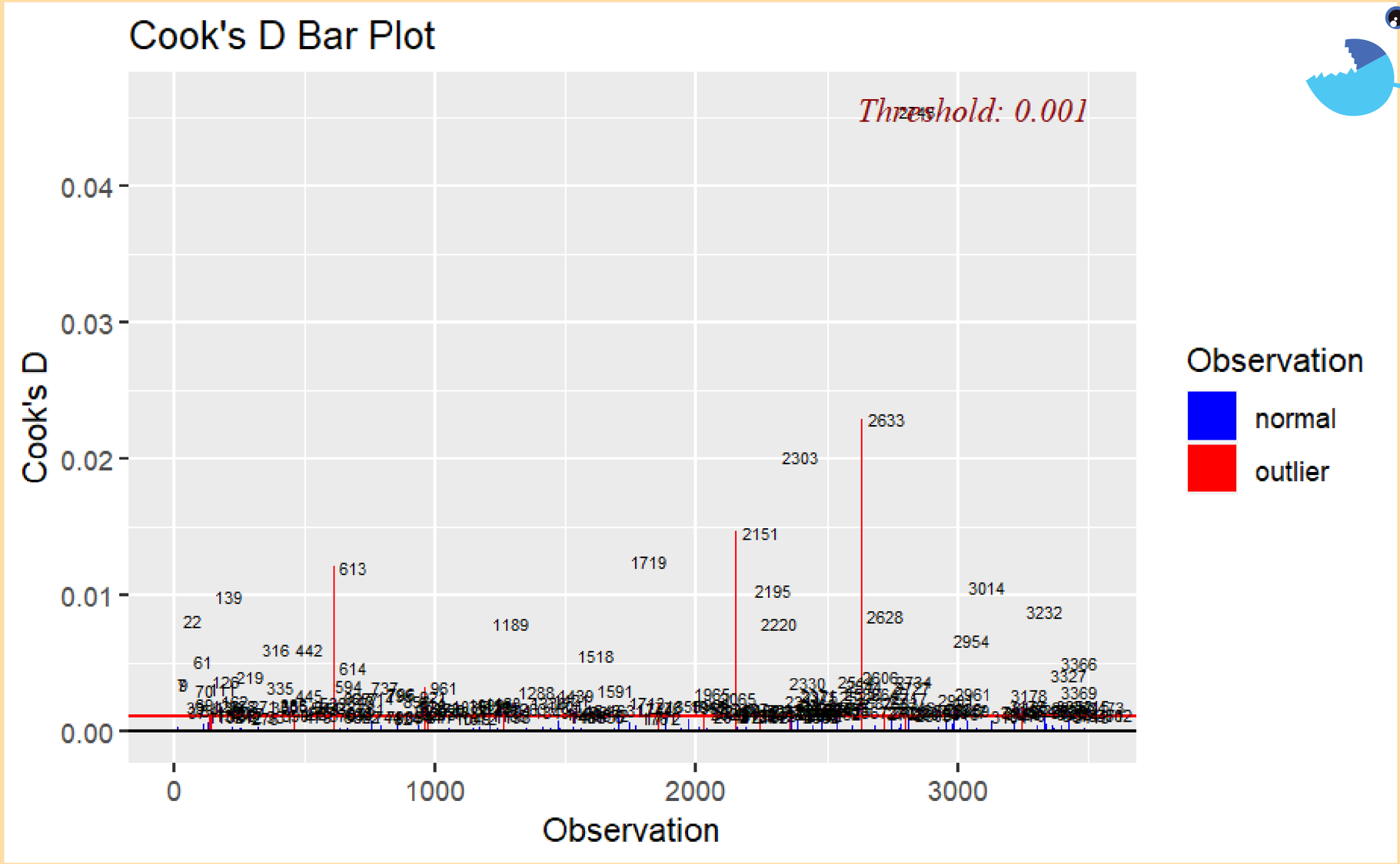
Studentized residuals are more effective for detecting outlying Y observations than standardized residuals.

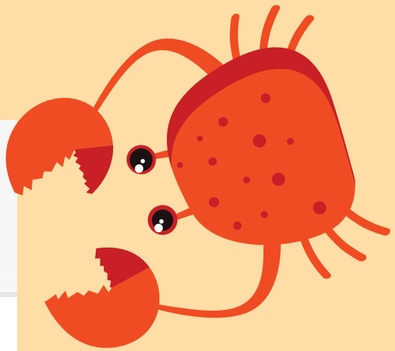




Observations with Studentized Residual larger than 3 (in absolute value) could be consider outliers.

A data point having a large Cook’s D indicates that the data point strongly influences the fitted values.

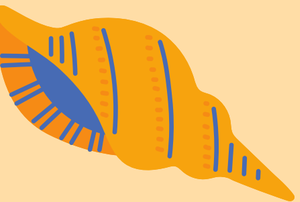




```{r}
sqrt(vif(mod))>5
```

```
          Length         Diameter           Height           Weight
            TRUE             TRUE            FALSE             TRUE
 Shucked.Weight Viscera.Weight     Shell.Weight            Sex_M
            TRUE            FALSE            FALSE            FALSE
           Sex_I
           FALSE
```

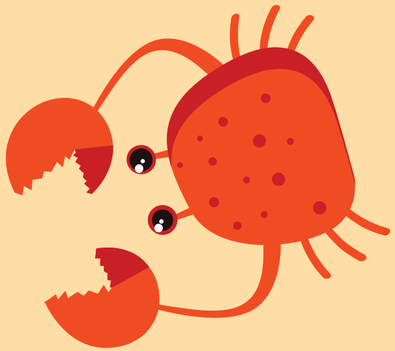
Some variables show multicollinearity:
to fix it we can drop some of them using subset selection or try Ridge and Lasso approaches.



SUBSET SELECTION

- **Identify a subset of relevant features that we believe to be related to the response.**

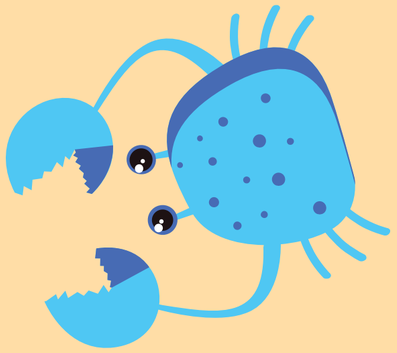
FORWARD STEPWISE SELECTION



- **Starting model: no predictors**
- **Adds predictors to the model, one-at-a-time**
- **Ending model: all the predictors are in the model**

Then select a single best model from among all using the values of Cp.

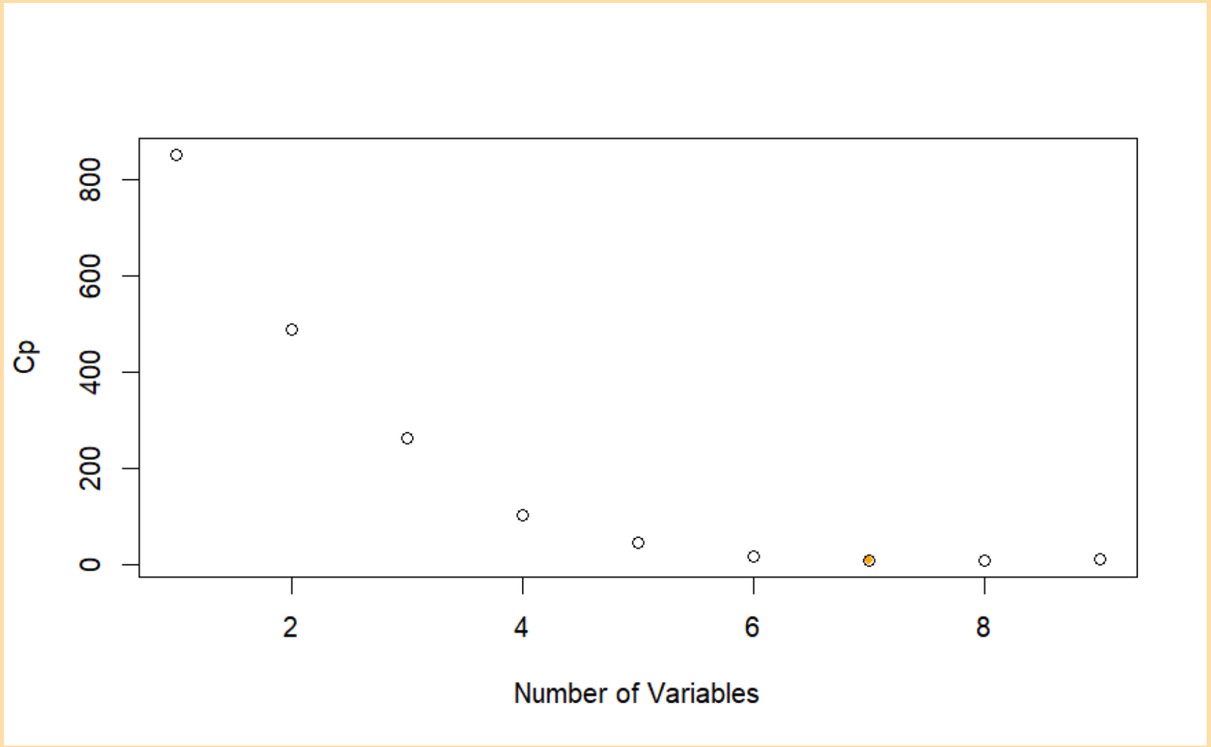
Call: regsubsets.formula(Age ~ . - Sex - Sex_F, data = ds1, nvmax = 9, method = "forward")

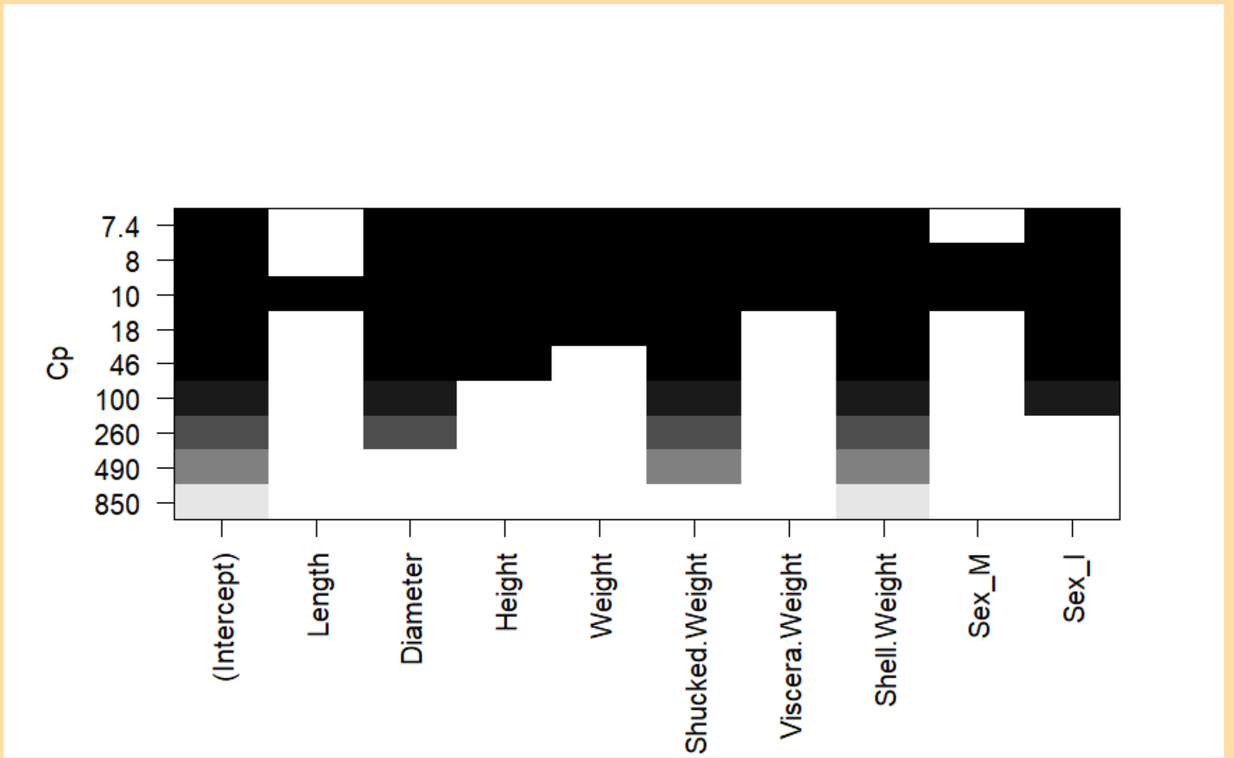


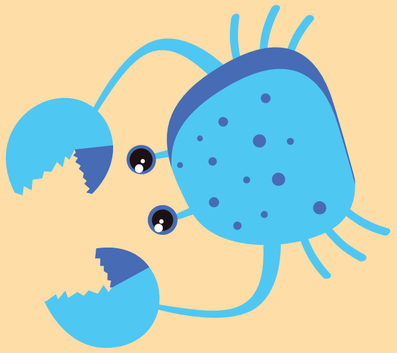
Selection Algorithm: forward

	Length	Diameter	Height	Weight	Shucked.Weight	Viscera.Weight	Shell.Weight	Sex_M	Sex_I
1	" "	" "	" "	" "	" "	" "	"*"	" "	" "
2	" "	" "	" "	" "	"*"	" "	"*"	" "	" "
3	" "	"*"	" "	" "	"*"	" "	"*"	" "	" "
4	" "	"*"	" "	" "	"*"	" "	"*"	" "	"*"
5	" "	"*"	"*"	" "	"*"	" "	"*"	" "	"*"
6	" "	"*"	"*"	"*"	"*"	" "	"*"	" "	"*"
7	" "	"*"	"*"	"*"	"*"	"*"	"*"	" "	"*"
8	" "	"*"	"*"	"*"	"*"	"*"	"*"	"*"	"*"
9	"*"	"*"	"*"	"*"	"*"	"*"	"*"	"*"	"*"

BEST SUBSET


Cp
0
200
400
600
800
2
4
6
8
Number of Variables


Cp
7.4
8
10
18
46
100
260
490
850
(Intercept)
Length
Diameter
Height
Weight
Shucked.Weight
Viscera.Weight
Shell.Weight
Sex_M
Sex_I

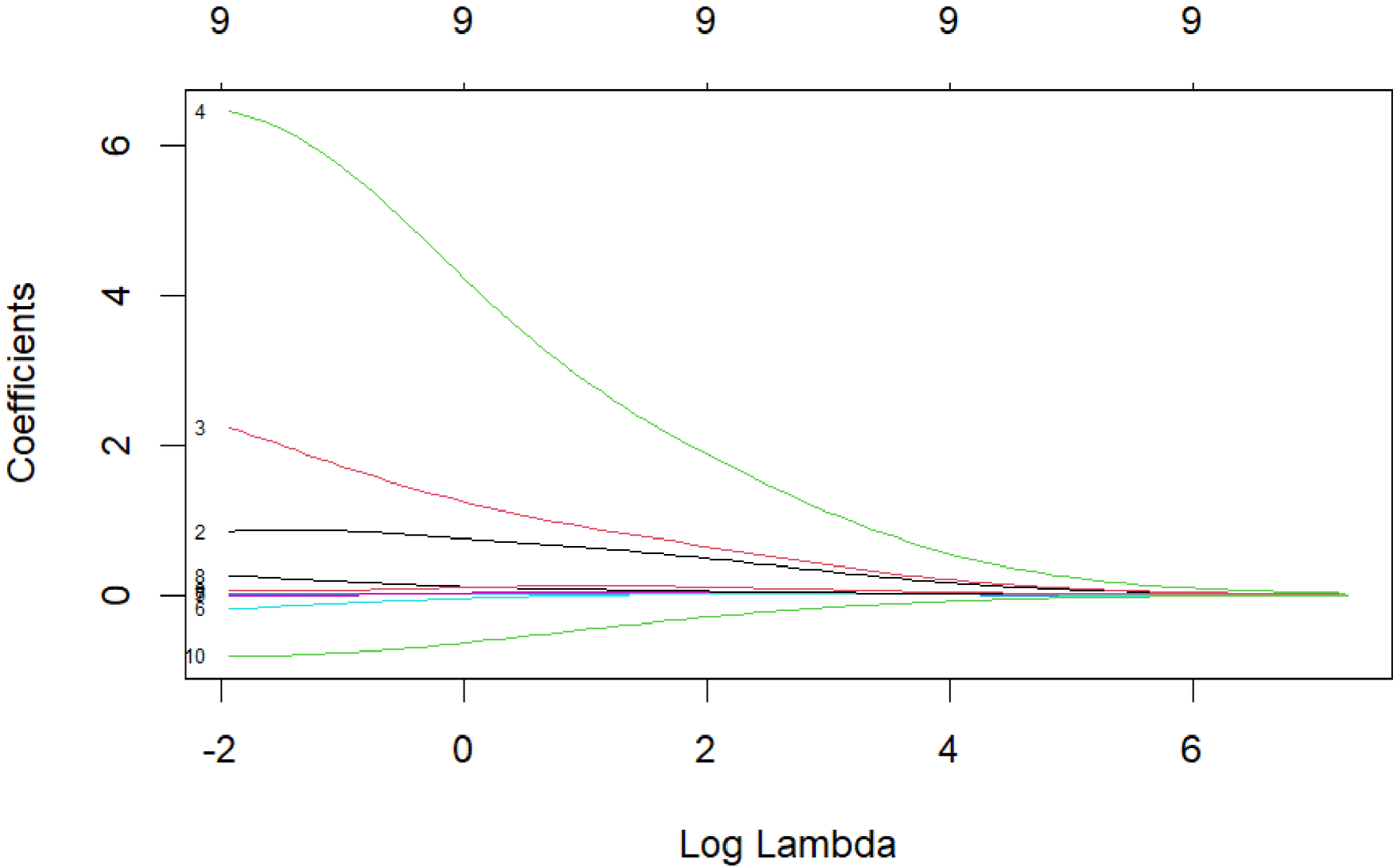


RIDGE REGRESSION

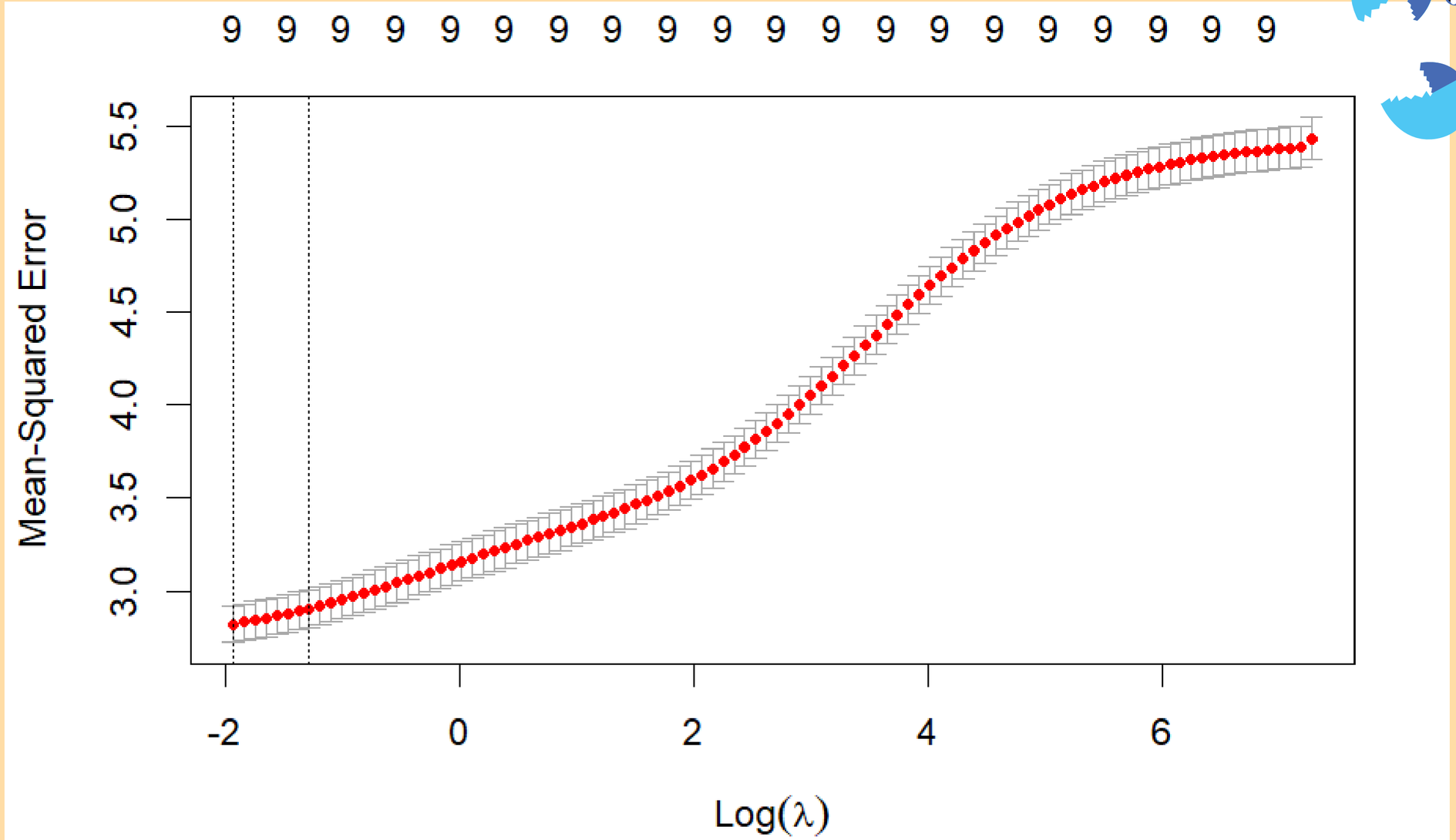
Let's try to use a Ridge Model to predict our Y

- **Fit of the model**
- **Cross Validation**
- **Best tuning parameter with train/validation division**
- **Coefficients and Accuracy**

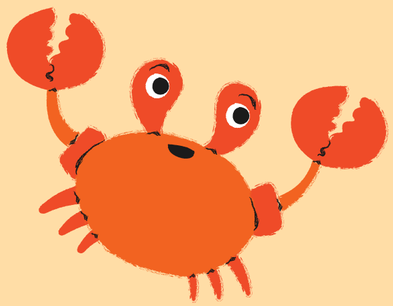
FIT OF THE MODEL



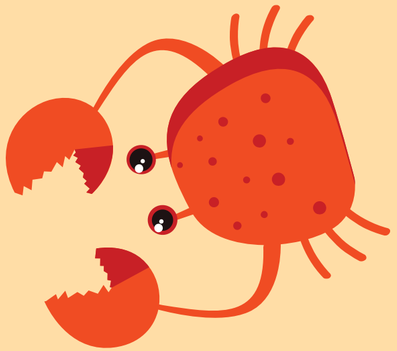
CROSS VALIDATION



We used the split in training and validation set to select the best value for lambda and compute the coefficients



COEFFICIENTS WITH BEST LAMBDA



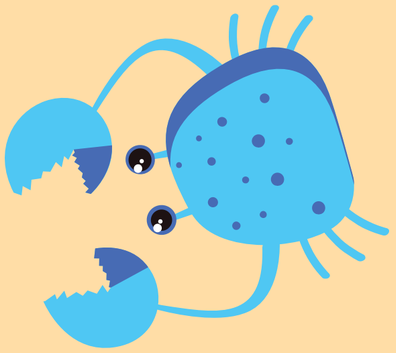
(Intercept)	4.64894979
Length	0.74958409
Diameter	2.66224804
Height	3.07153969
Weight	0.03166028
Shucked.Weight	-0.27049230
Viscera.Weight	-0.06120432
Shell.Weight	0.46660092
Sex_M	0.02041585
Sex_I	-1.00133582

MSE with best lambda = 2.783479

R-squared = 0.4796713

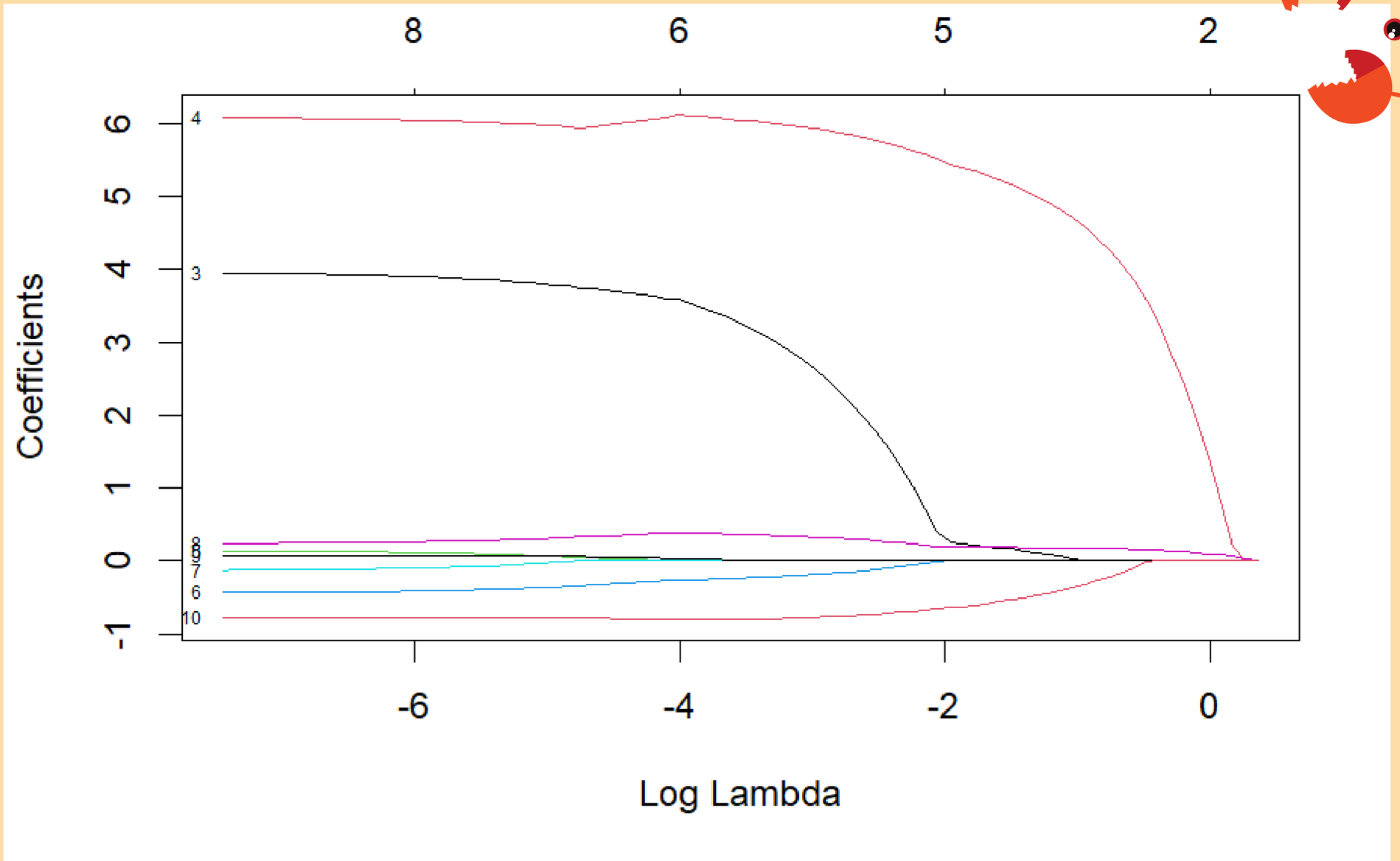
No improvement in respect to linear regression

THE LASSO

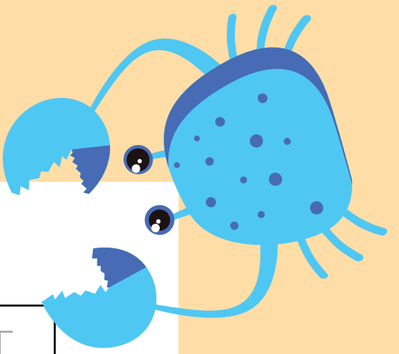


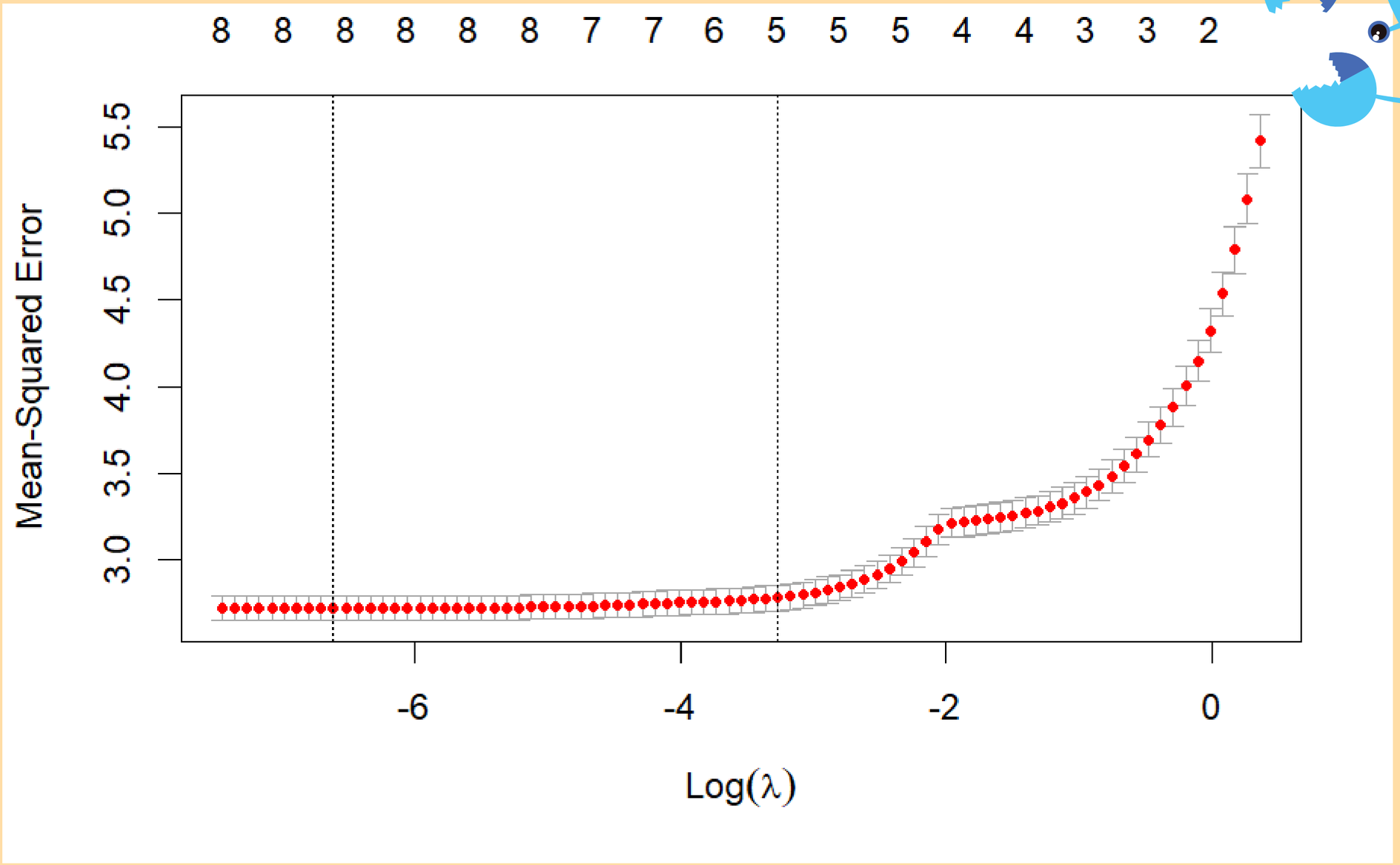
- Fit of the model
- Cross Validation
- Best tuning parameter with train/validation division
- Coefficients and Accuracy

FIT OF THE MODEL

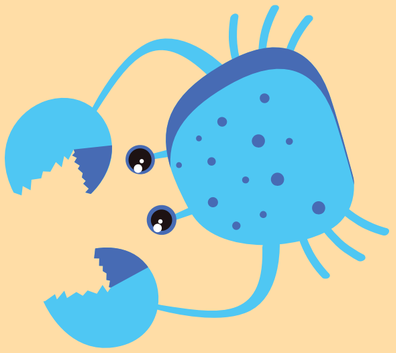


CROSS VALIDATION





COEFFICIENTS WITH BEST LAMBDA



(Intercept)	3.98850578
Length	-0.69873939
Diameter	5.02768432
Height	5.59732819
Weight	0.11178246
Shucked.Weight	-0.42639457
Viscera.Weight	-0.08175656
Shell.Weight	0.27098462
Sex_M	0.05594342
Sex_I	-0.88895530

MSE with best lambda = 2.807608

R-squared =0.5018999

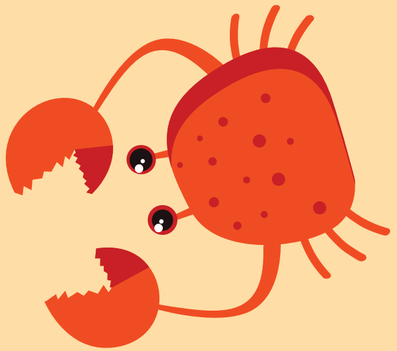
No significative improvement in respect to linear regression and almost the same MSE of the Ridge

RANDOM FOREST

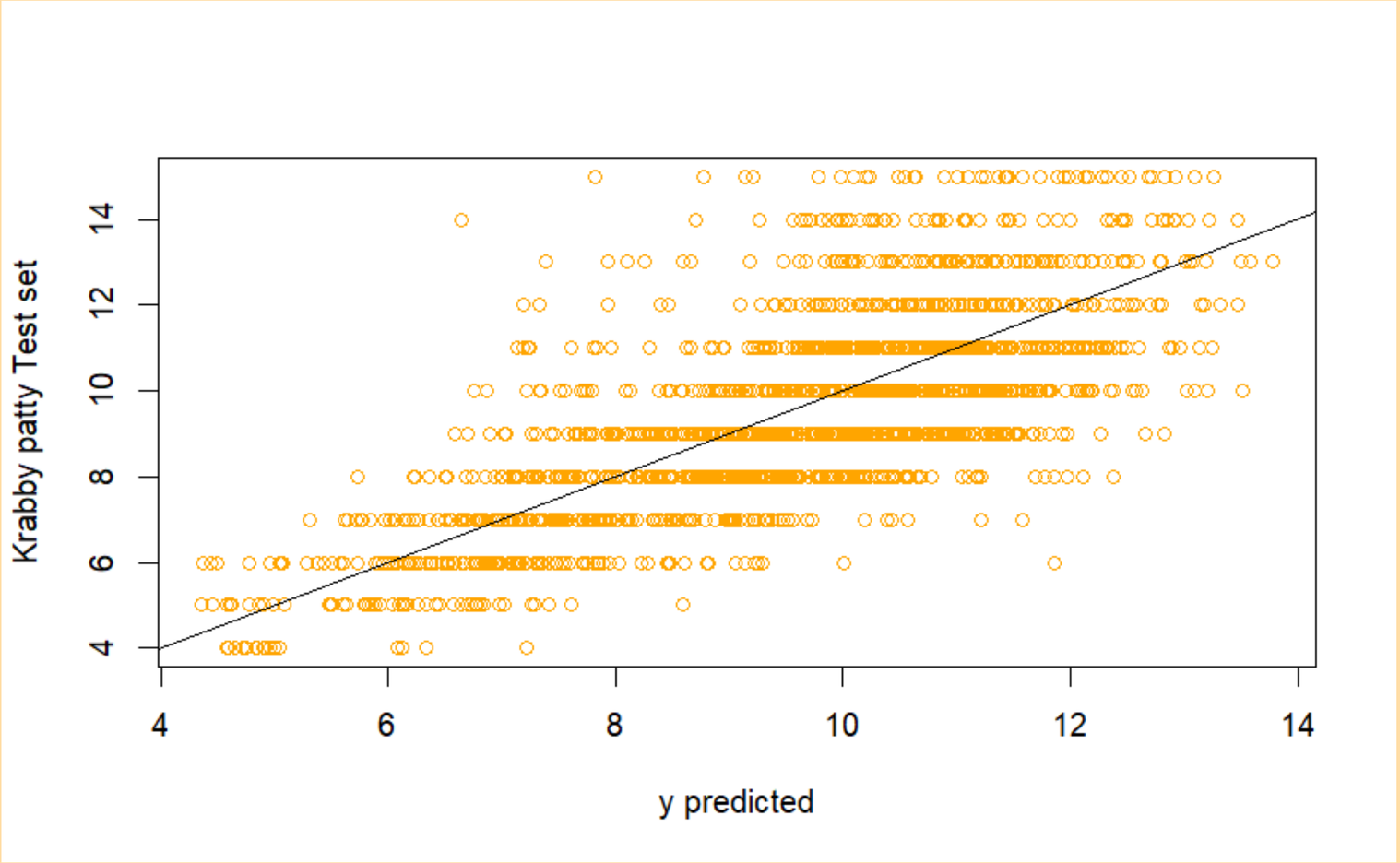
Divide the data into training and test set

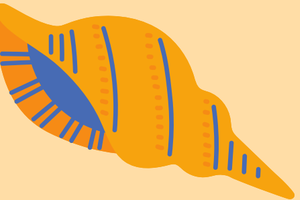
Output: average prediction of the trees

The root mean square error (RMSE) measures the average difference between a statistical model's predicted values and the actual values

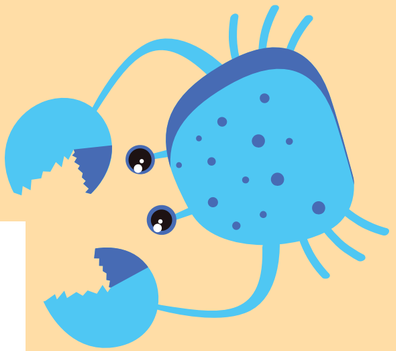


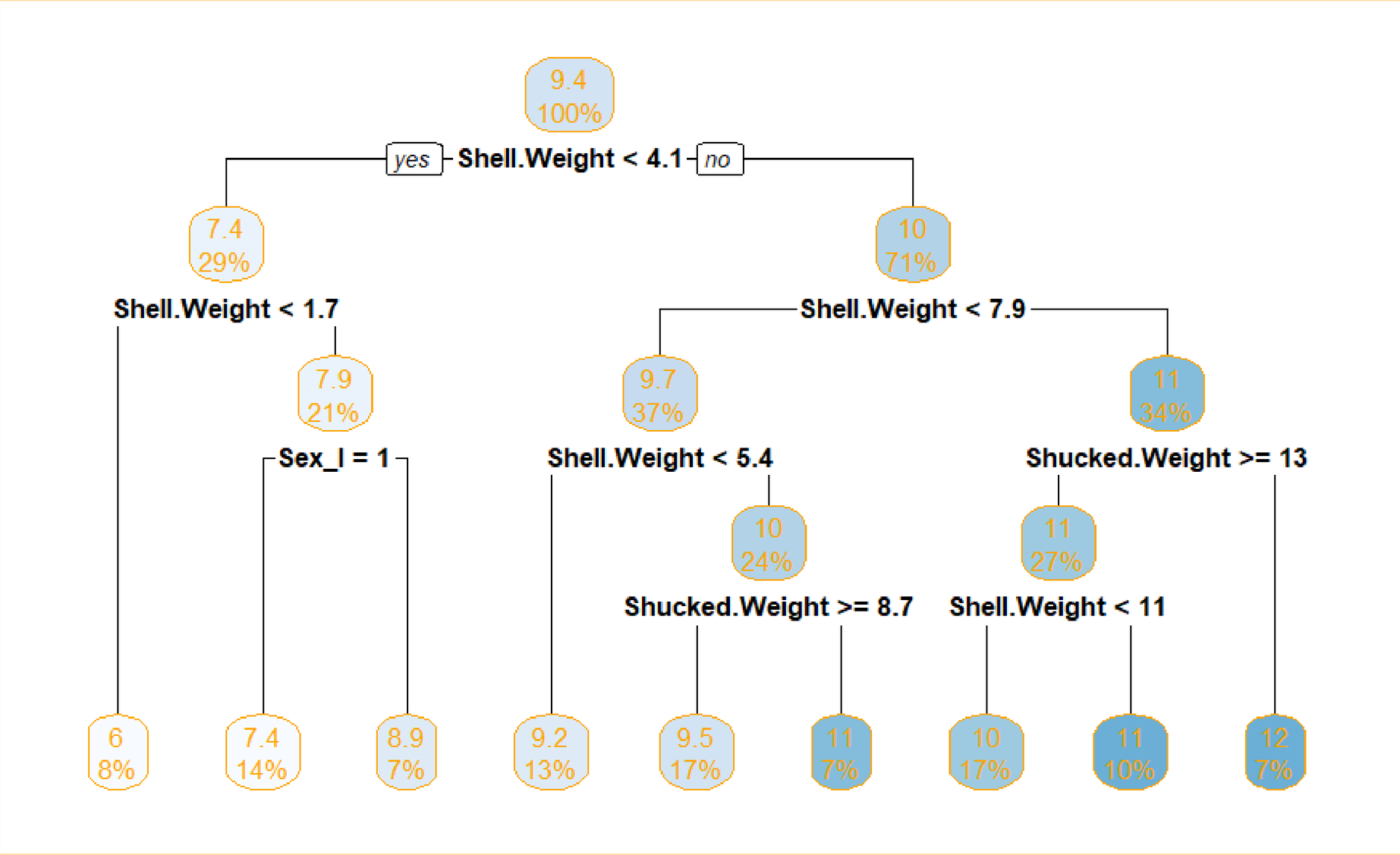
$$\sqrt{MSE} = 1.226756$$



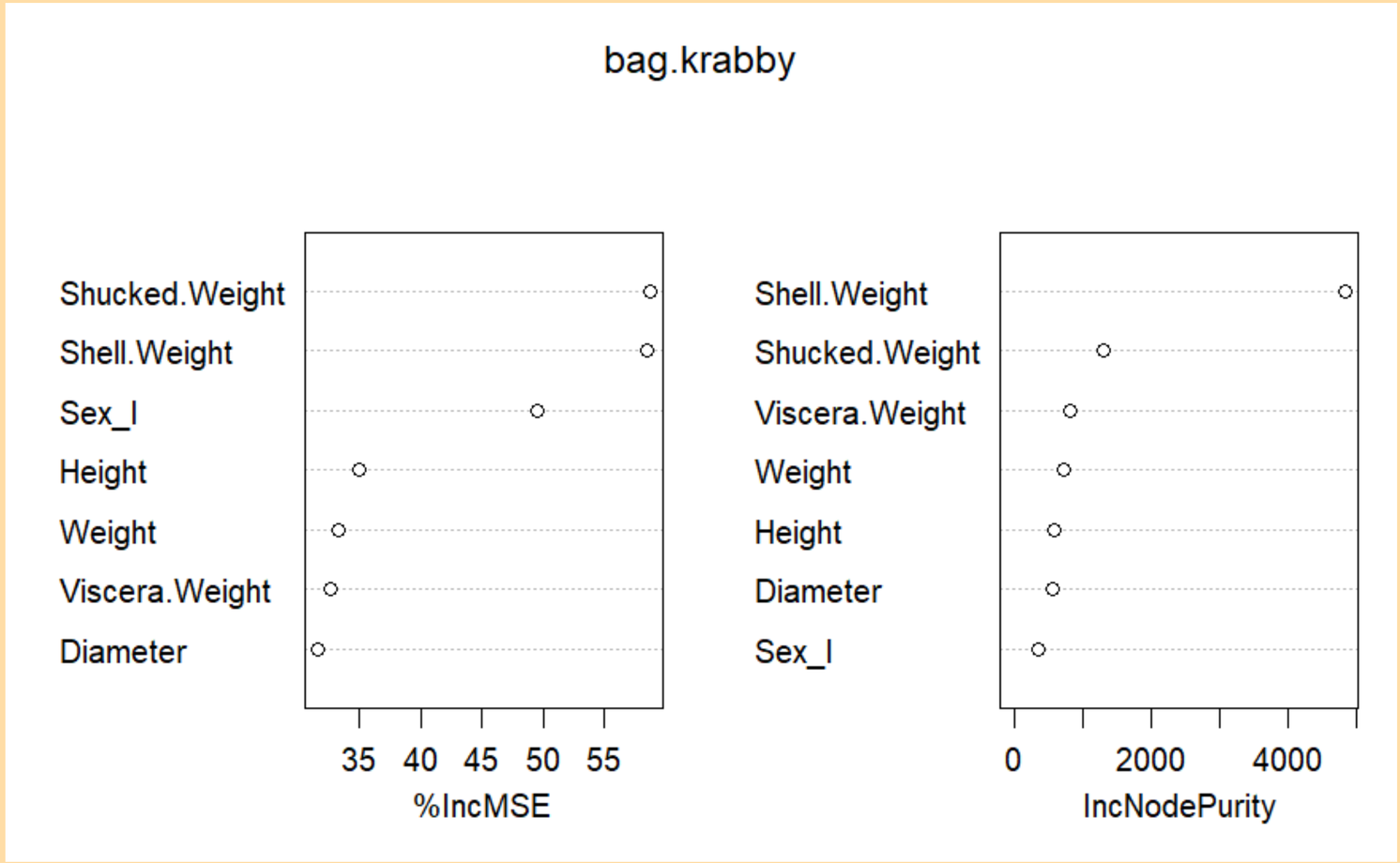


TREE RAPPRESENTATION

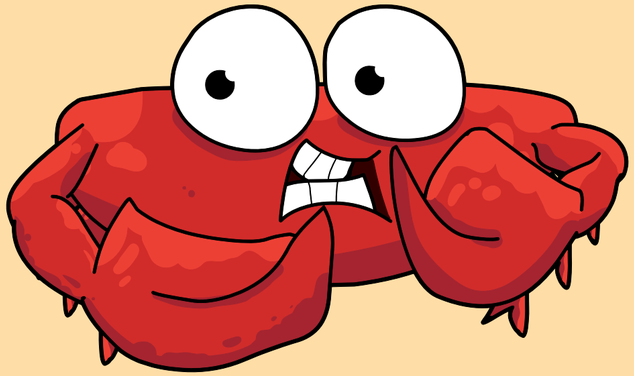




Most important variables



UNSUPERVISED LEARNING

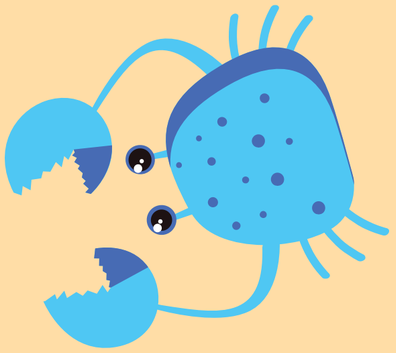


PCA

- Goal: reduce the dataset and find a suitable number of components explaining the majority of the variance in the dataset

- We want to show how the samples are related (or not related) to each other

THE PCA OUTPUTS

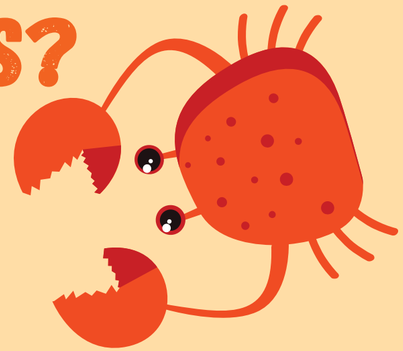


- ***The Standard Deviation***

- ***The Rotation matrix***: provides the Principal Component loading vectors

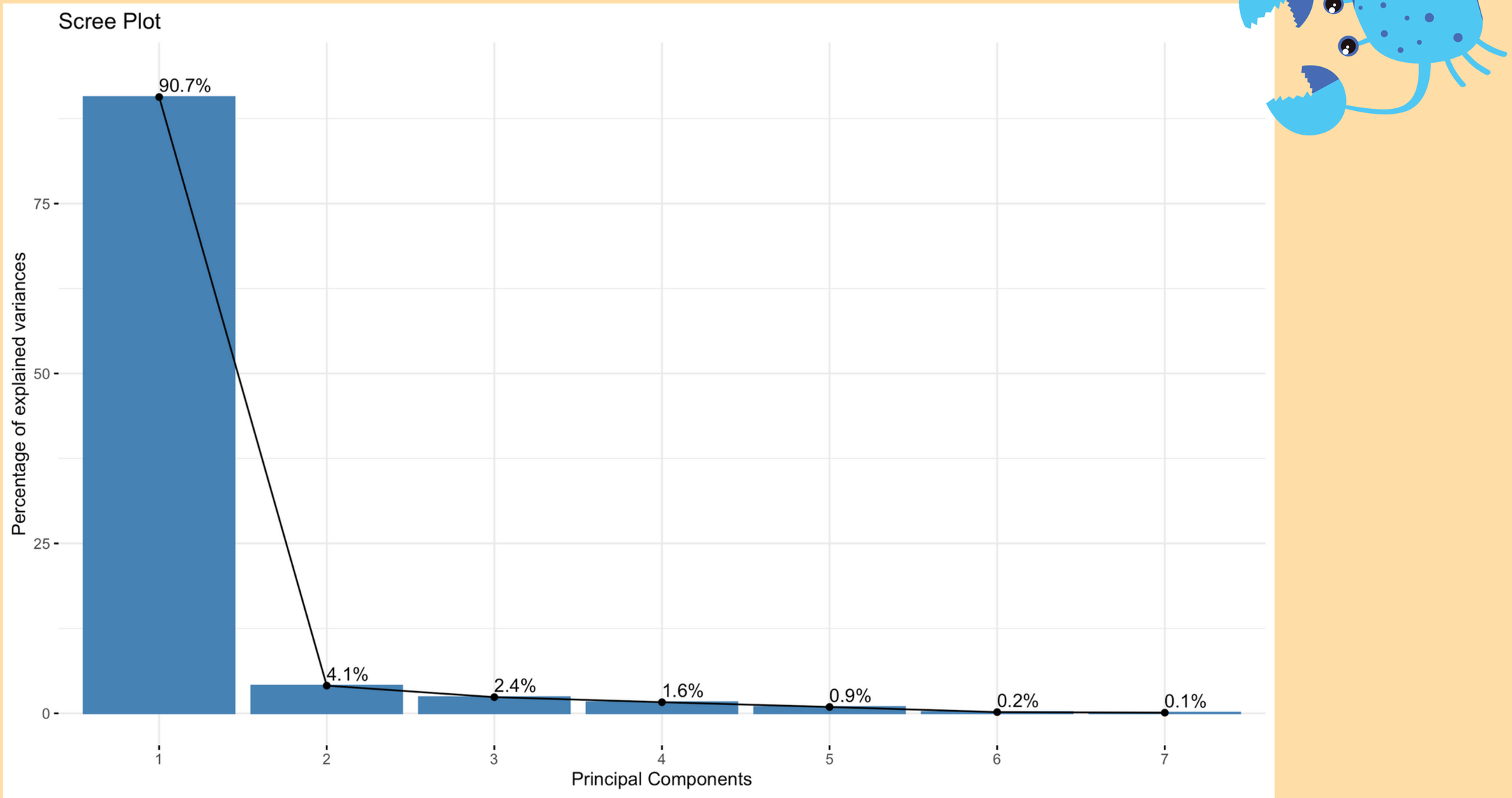
- ***x***: provides the Principal Component score vectors

WHAT ARE THE PRINCIPAL COMPONENTS?

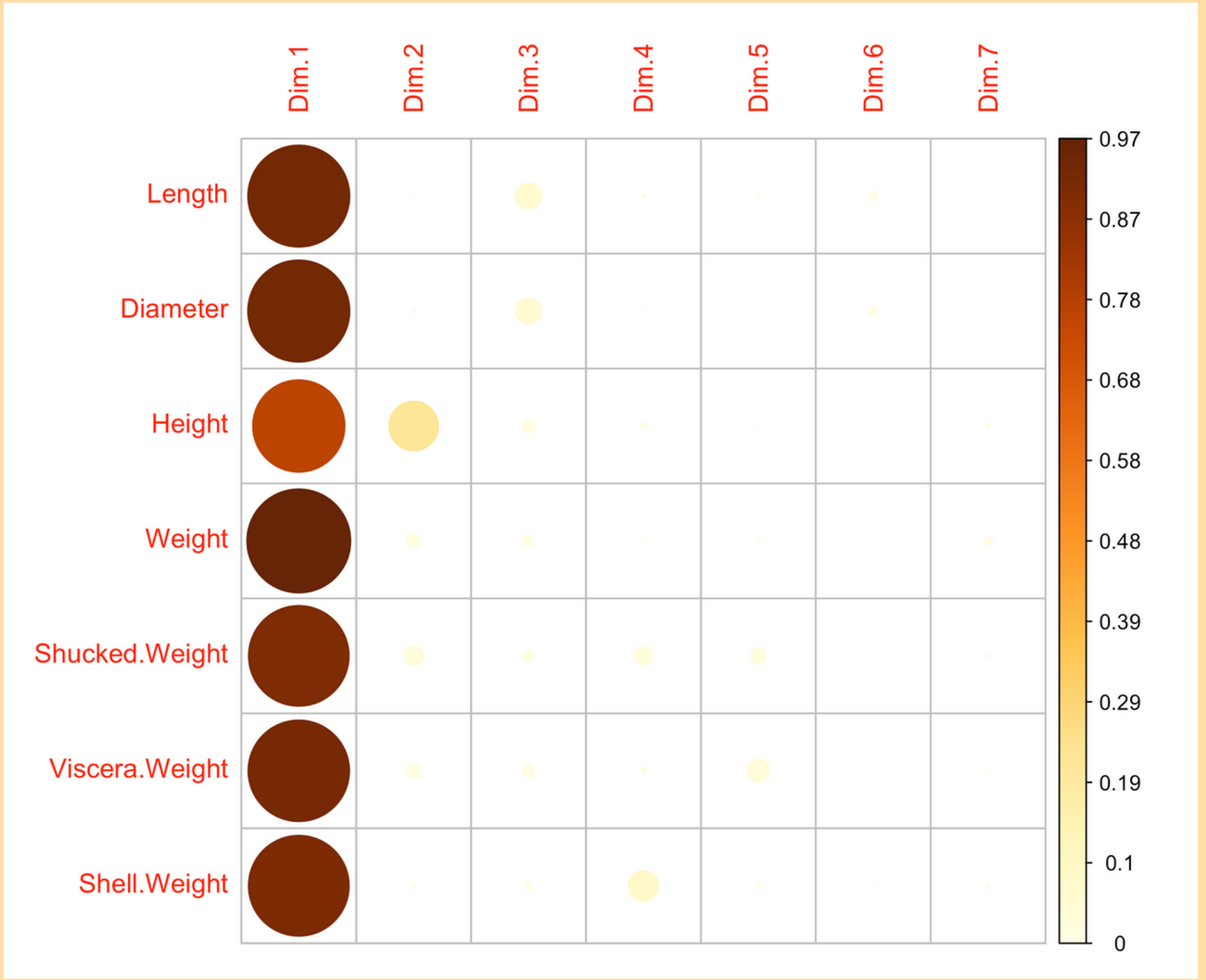


- The PC is a new variable
- The PCs are uncorrelated
- Most of the information of the original variables are compressed in the first component
- The second most information are compressed in the second component and so on up until we have the Scree Plot

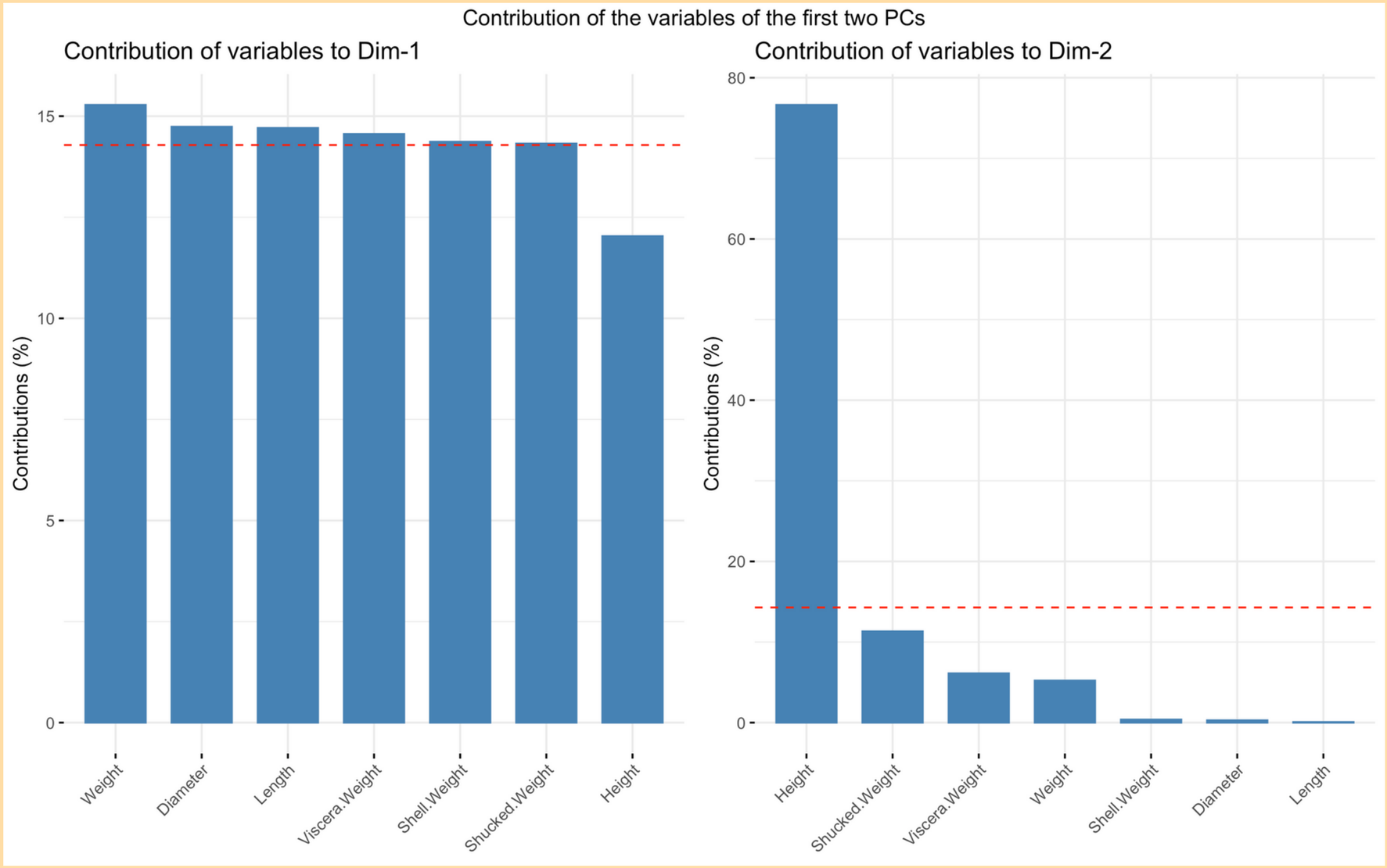
SCREE PLOT


Scree Plot
90.7%
4.1%
2.4%
1.6%
0.9%
0.2%
0.1%
Percentage of explained variances
0
25
50
75
1
2
3
4
5
6
7
Principal Components

We can extract the PC result for variables only



In this way we can see the most contributing variables for each dimension

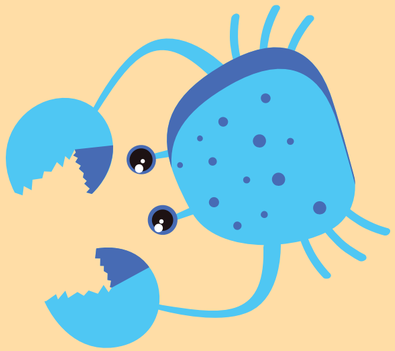


In the graph we can see the comparison of the contribution of the variables in the first two Principal components

CLUSTERING

- finding subgroups or clusters
- goal <- partition of the data into distinct groups that share similarities within them
- Two common methods: *k-means clustering* and *hierarchical clustering*

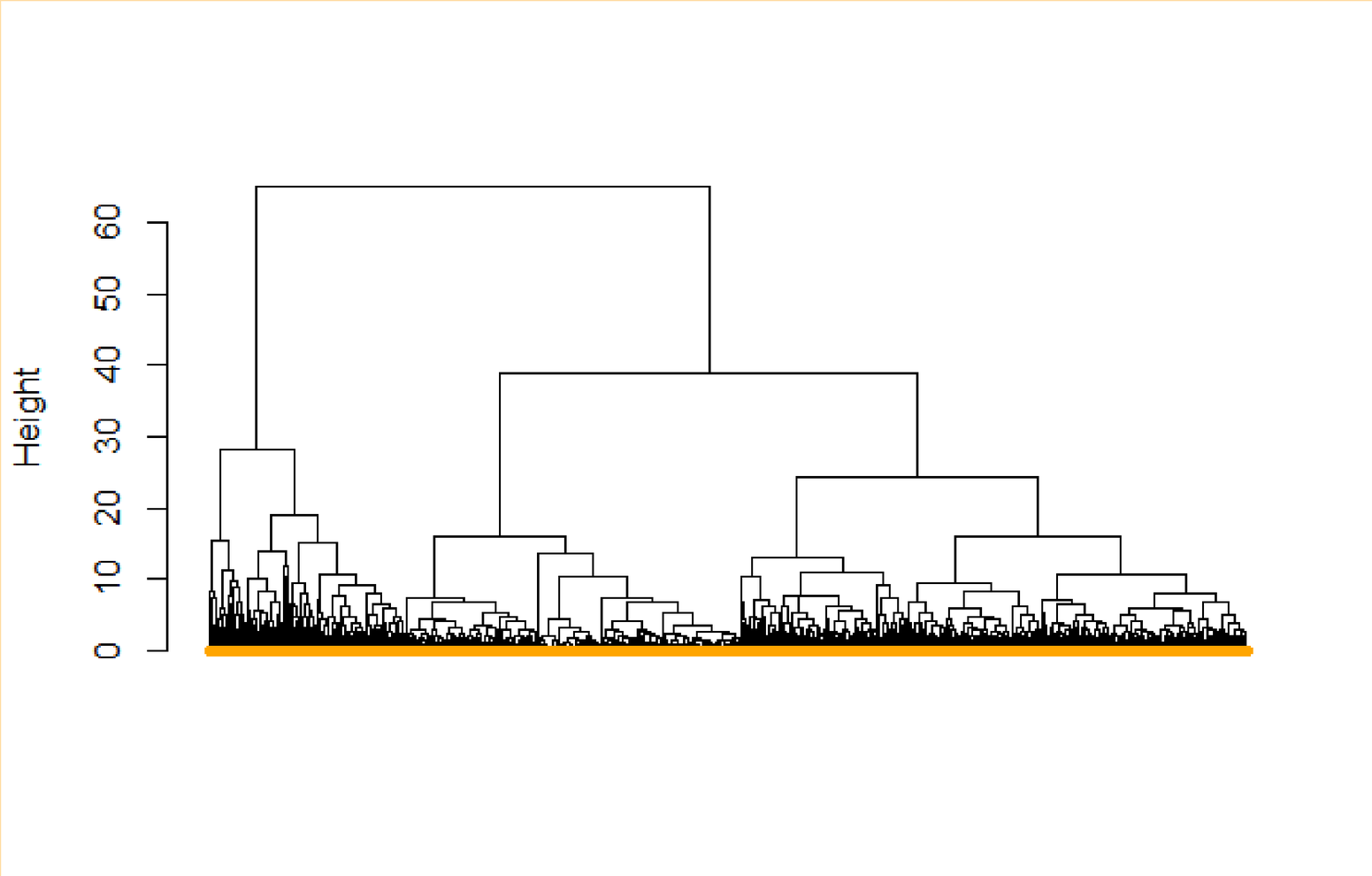
HIERARCHICAL CLUSTERING



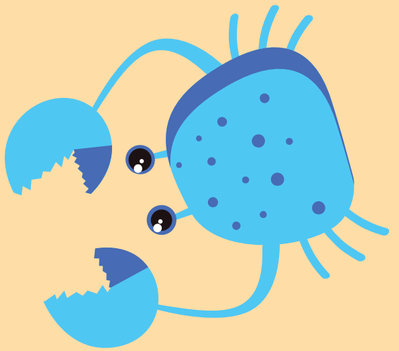
Hierarchical clustering, unlike K-Means Clustering, does not require a pre-selected number of cluster.

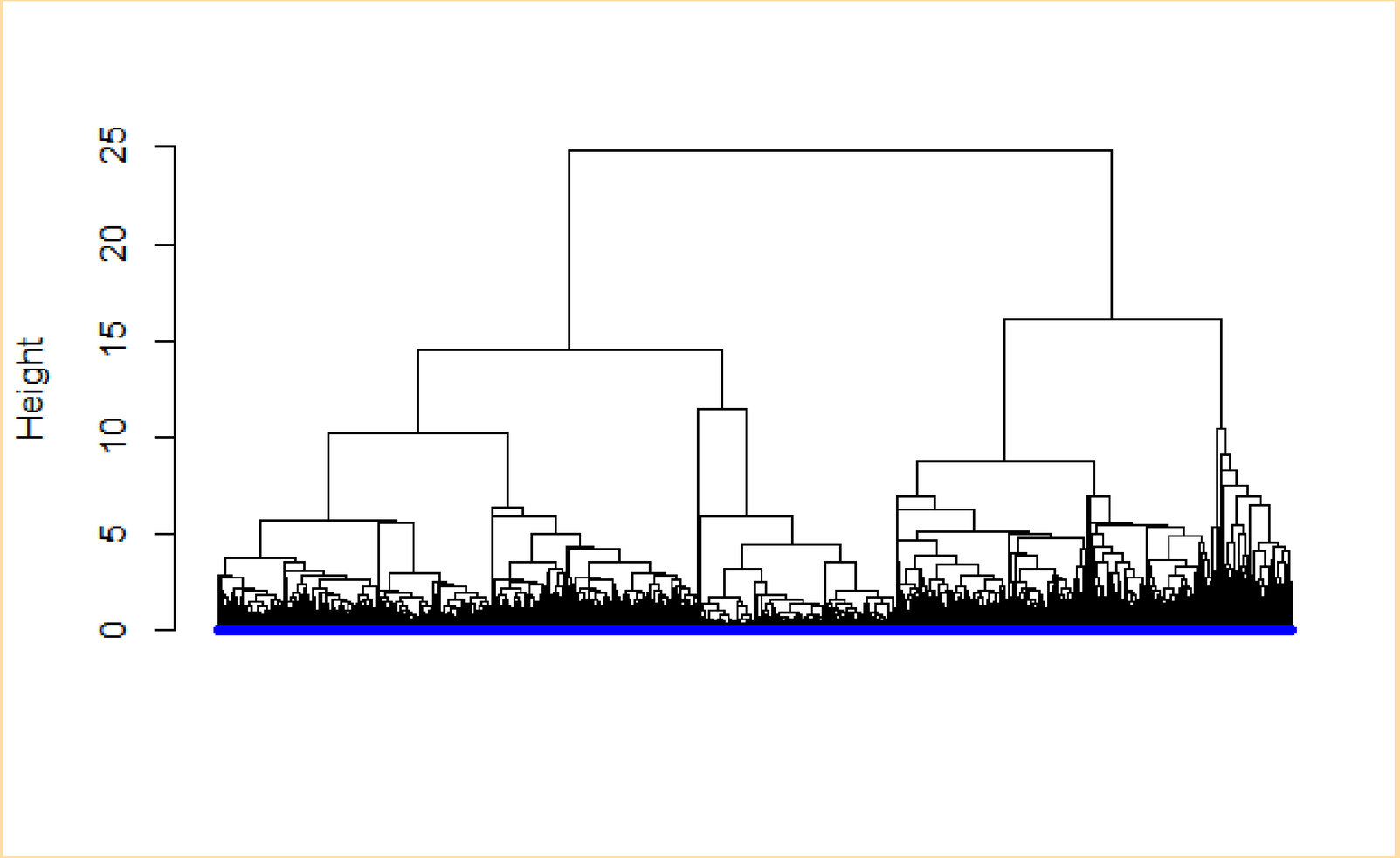
It can be a useful tool to detect patterns in your data even when you do not have a designated outcome variable

COMPLETE LINKAGE DENDOGRAM

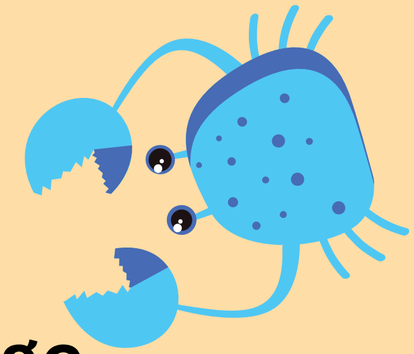


AVERAGE LINKAGE DENDOGRAM

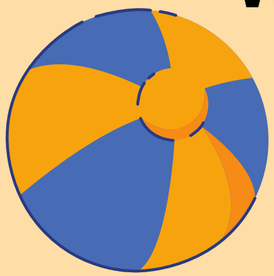




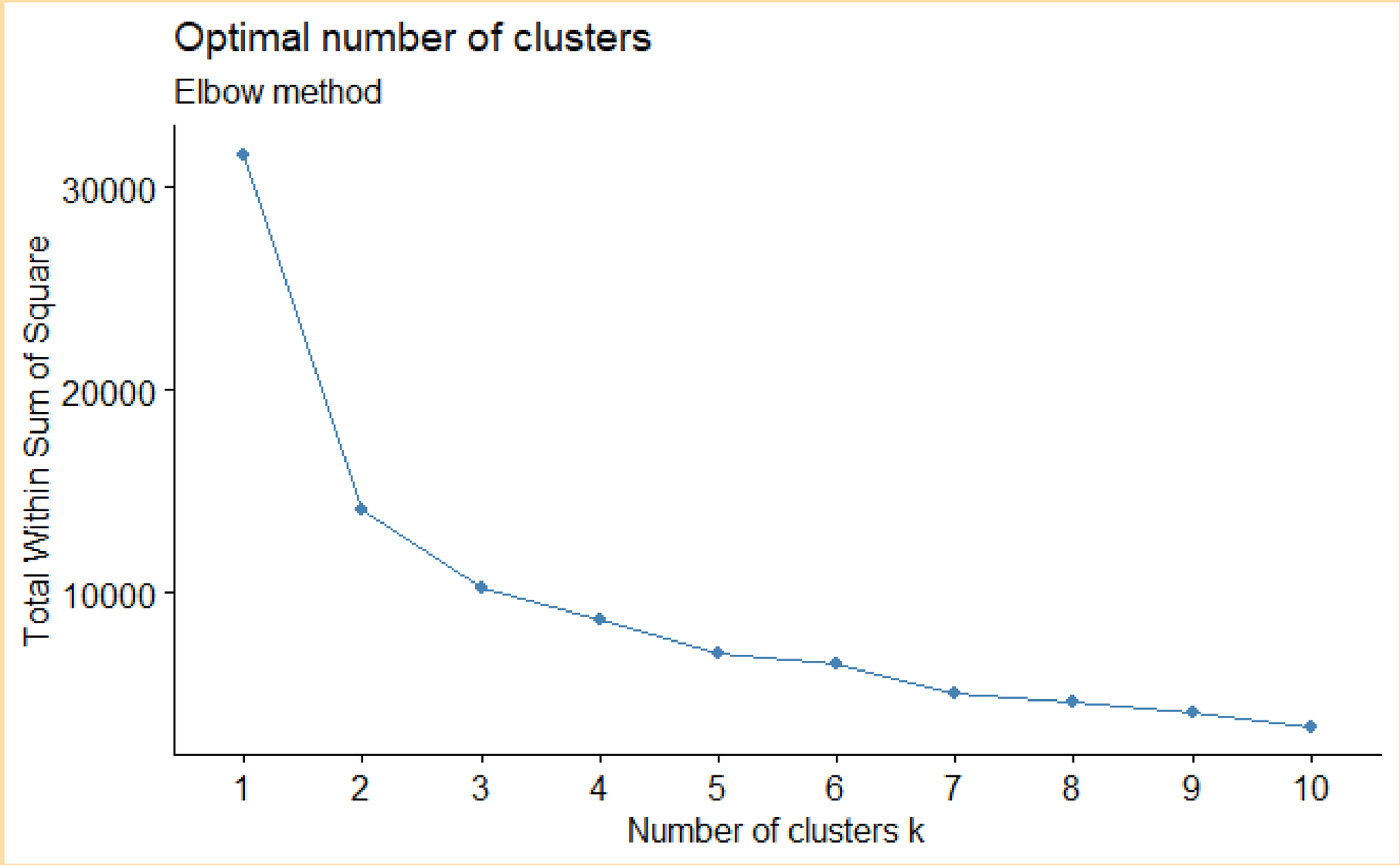
K-MEANS



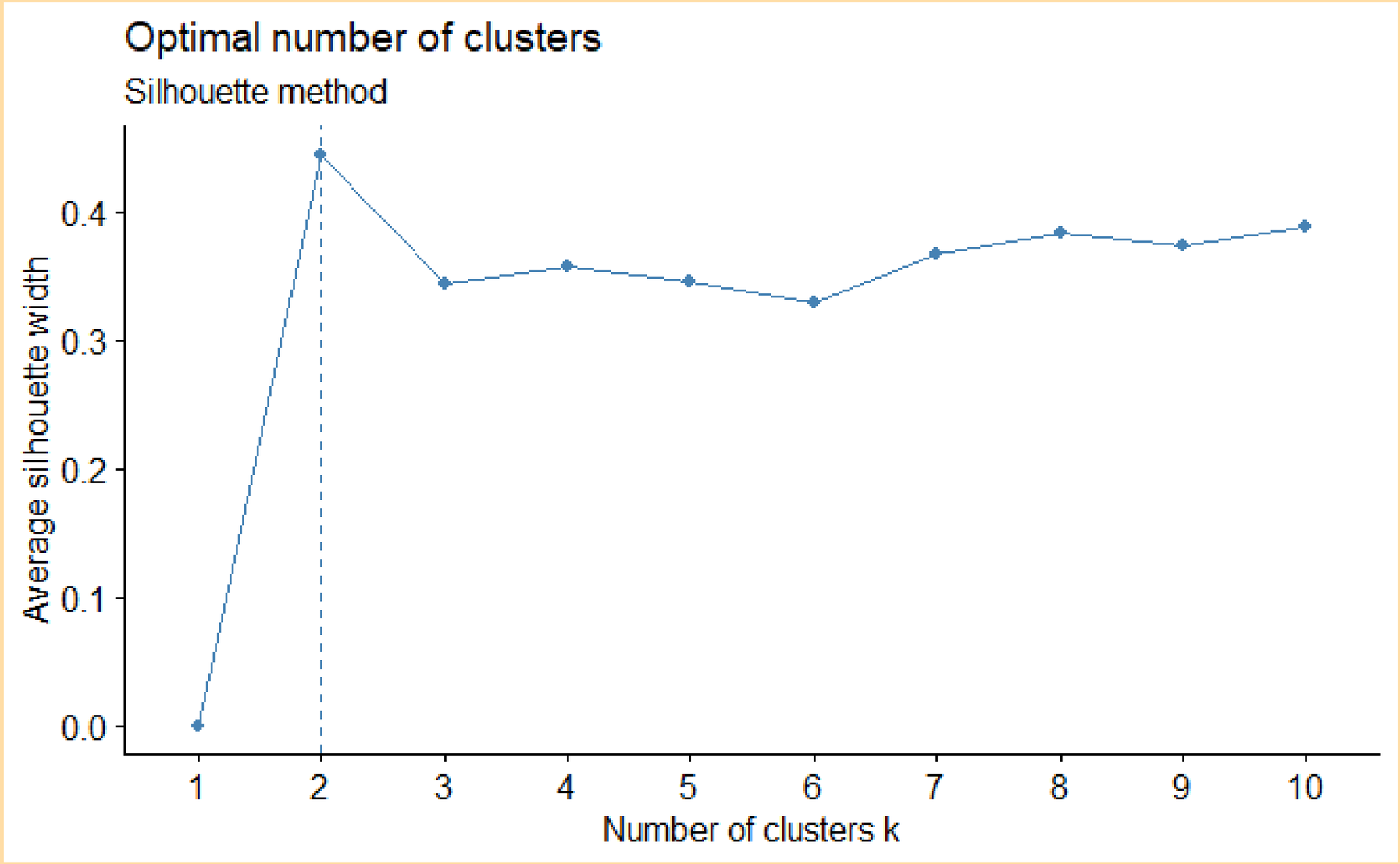
- Our features: Length, Diameter, Weight, Height, Shucked Weight, Viscera Weight, Shell Weight, Age, Sex_F, Sex_I, Sex_M
- We removed Age since we wanted to predict it
- We standardised our data using the scale function
- We looked for the optimal number of clusters

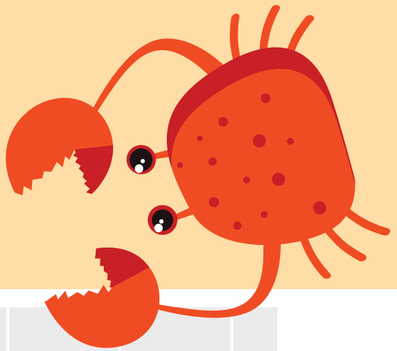


THE ELBOW METHOD



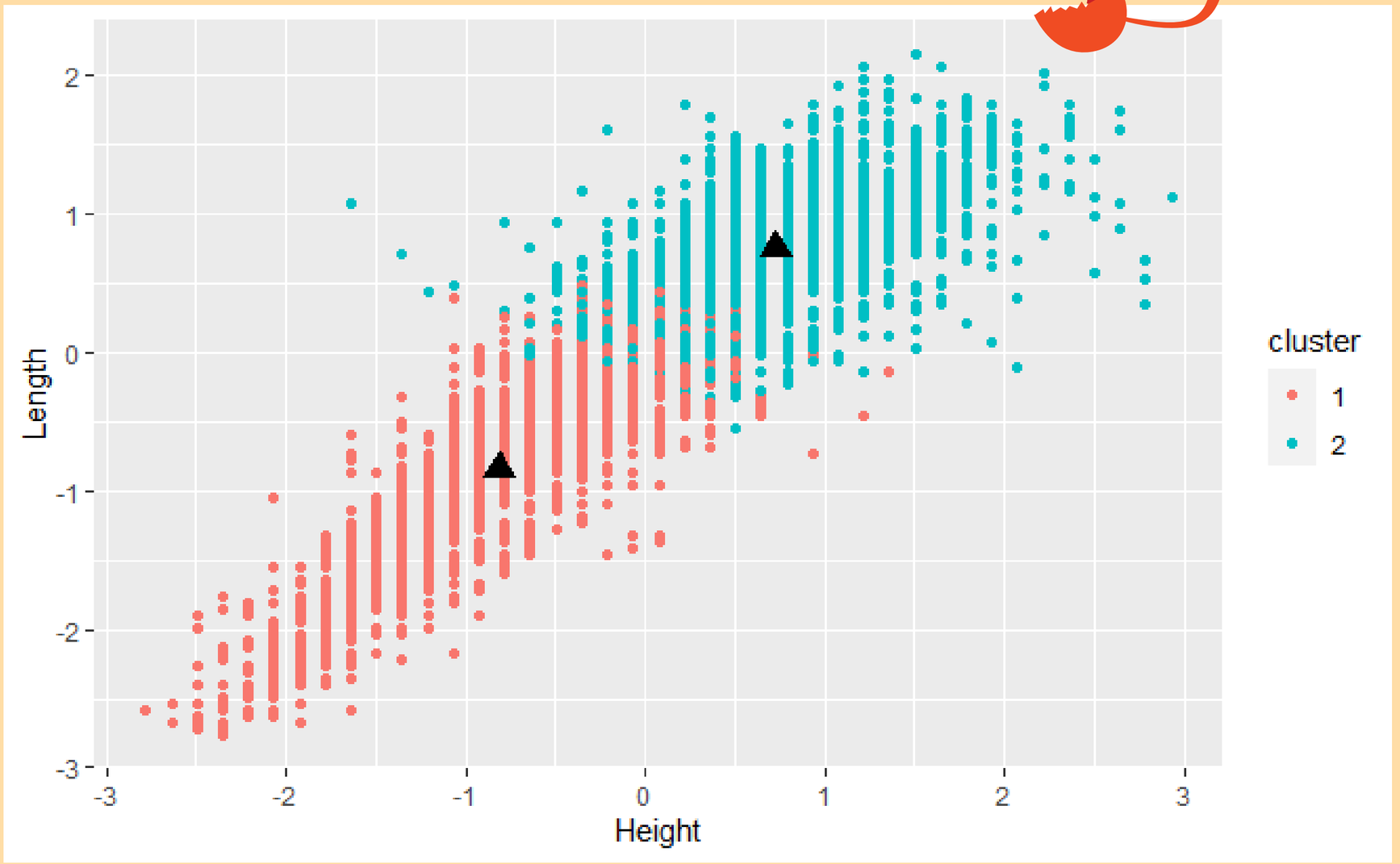
THE SILHOUETTE METHOD





- **2 clusters were selected based on the previous tests (2 out of 2 tests)**

- **We get the clusters results**



Here we have the clusters according to Length and Weight of the crabs

We can clearly notice that there is a mixing of the two clusters in the middle

EVALUATION AND INTERPRETATION

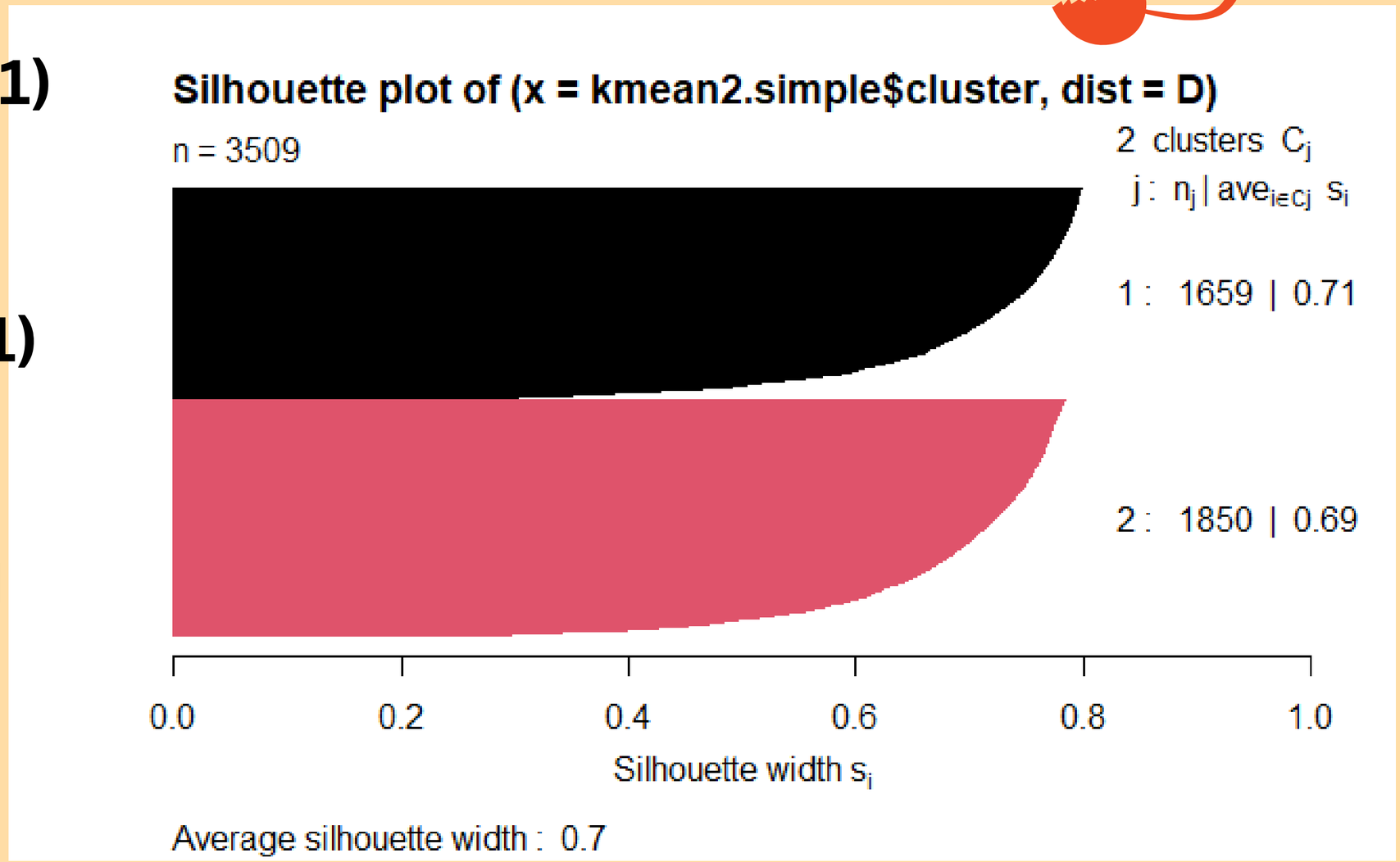
- **Cluster 1: 0.71**

(Good as it is close to 1)

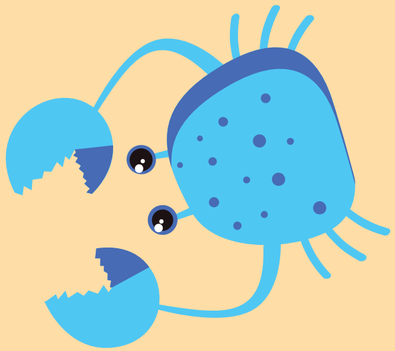
- **Cluster 2: 0.69**

(Good as it is close to 1)

The internal cluster quality evaluation is good as shown in the plot.



LAST FINDINGS...



- 1659 and 1850 observations respectively. A ration of 47% to 52%.
- Cluster 1 is higher than cluster 2 based on the centroid mean.
- Intra cluster bond strength factor Cluster 1: 6578.707 Cluster 2: 7547.571
- Goodness of the classification k-means: 55.3% (slightly good fit)
- Between Clusters: 17445.72

SO, WHAT IS OUR FINAL RESULT?

MR KRAB

Sex	Length	Diameter	Height	Weight	Shucked.Weight	Viscera.Weight	Shell.Weight
M	1.5250	1.1750	0.3875	29.270859	14.089702	6.1660163	7.8953358

RANDOM FOREST PREDICTION

predict(bag.krabby, MrKrab, type="response", norm.votes=TRUE, predict.all=FALSE, proximity=FALSE, nodes=FALSE)

PREDICTED AGE

9.877333

Thank You!