Pin: 84169

1. Data preparation

```
spc_tbl_ [1,329 × 8] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
 $ product_id     : chr [1:1329] "0001" "0002" "0003" "0004" ...
 $ category       : Factor w/ 6 levels "Accessory","Equipment",..: 3 4 3 5 4 4 4 6 2 5 ...
 $ animal         : Factor w/ 4 levels "Bird","Cat","Dog",..: 1 1 3 2 2 3 3 2 4 3 ...
 $ size           : Factor w/ 3 levels "large","medium",..: 1 2 2 3 3 3 3 2 2 2 ...
 $ price          : num [1:1329] 51.1 36 31.2 24.9 26.2 ...
 $ sales          : num [1:1329] 1861 964 898 982 833 ...
 $ rating         : num [1:1329] 7 6 5 6 7 7 5 4 5 8 ...
 $ repeat_purchase: Factor w/ 2 levels "0","1": 2 1 2 2 2 1 1 1 1 1 ...
 - attr(*, "spec")=
  .. cols(
  ..   product_id = col_character(),
  ..   category = col_character(),
  ..   animal = col_character(),
  ..   size = col_character(),
  ..   price = col_double(),
  ..   sales = col_double(),
  ..   rating = col_double(),
  ..   repeat_purchase = col_double()
  .. )
 - attr(*, "problems")=<externalptr>
```

2. Exploratory Data Analysis (EDA)

2.1 Sales Value by Animal Segment

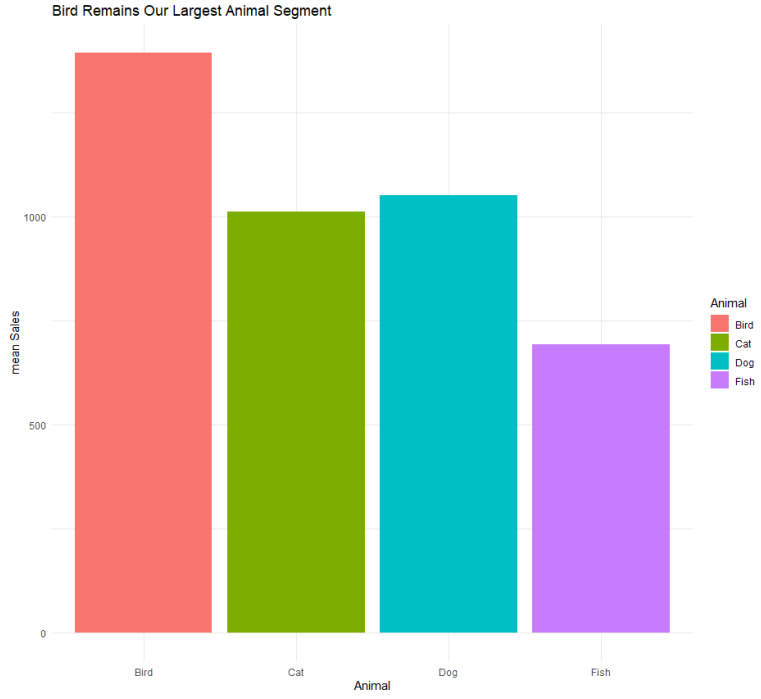


Figure 2.1 – Mean sales by Animal

2.1a) The Bird category generates the highest average sales value per product with a mean of 1393.62.

2.1b) The manager states about the bid category that “they earn the most”. Mean sales only capture sales per product. If any category had more sales they could earn Northpaw Pet Supplies more

overall. Furthermore, the sales does not capture costs. Where Birds could earn the most, but have excessive costs leading to loss making.

2.2 Price by animal segment

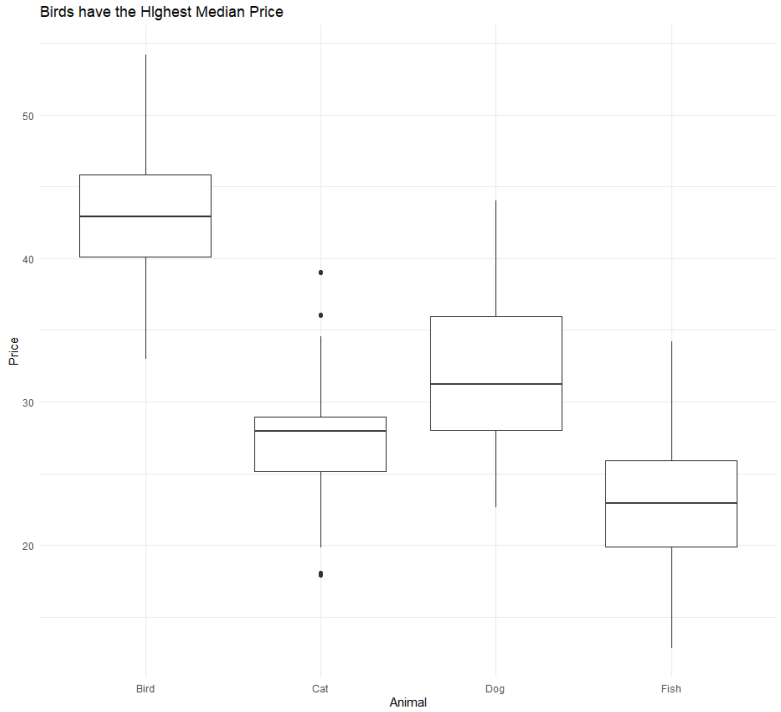


Figure 2.w – Boxplot of price by Animal

2.2a) The median of both the bird and fish category lies in the middle of the interquartile (IQ) range, suggesting that the median is very similar to the mean.

2.2b) the higher average sales value of the bird category in question 2.1 is explained by the fact that most observations for the price within the birds category is higher than those values observed in the other categories.

2.3 Building a Volume Measure

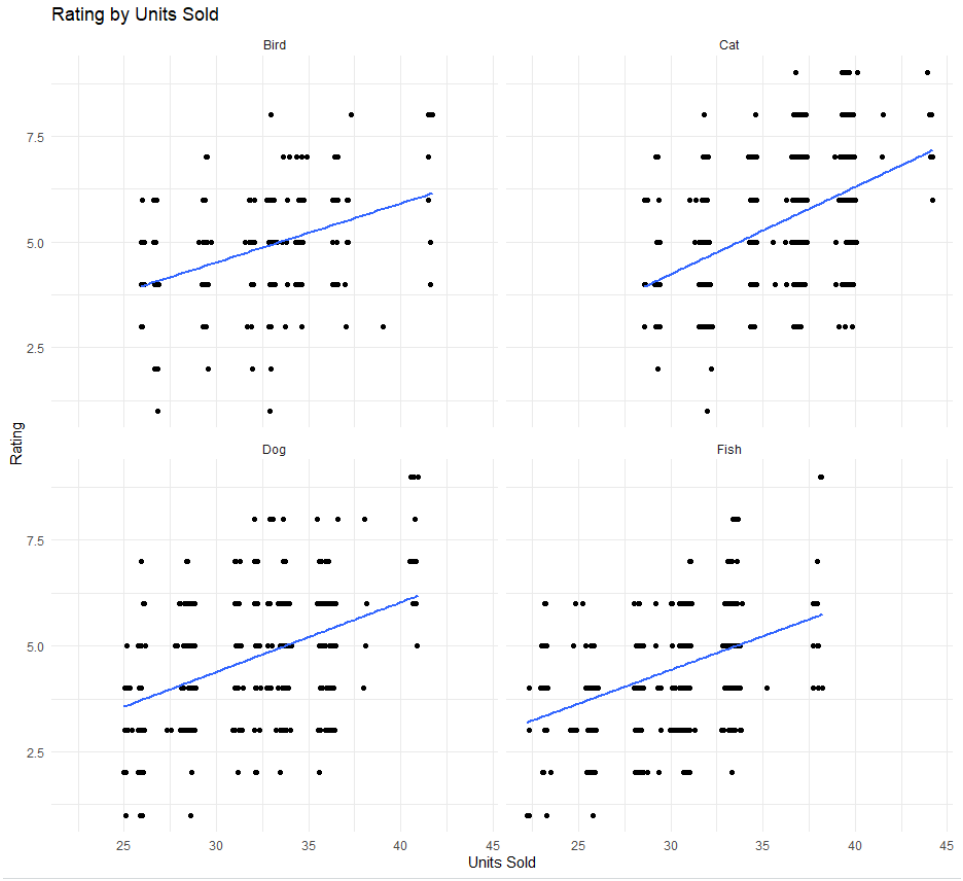


Figure 2.3 – Rating by Units Sold with regression line

2.3a) from the scatterplot we see that the cat segment sells more units on average. Section 2 analyzed the mean sales price only accounting for sales, while the volumetric aspect of total sales were not captured.

2.3b) The relationship between units sold and rating is positive in all categories. This could be explained by customers most likely buying products they see others have been satisfied with.

2.3c) The scatterplot and regression tells the retailer to focus on selling and marketing products with higher ratings as this increases sales.

2.4 Share of Business

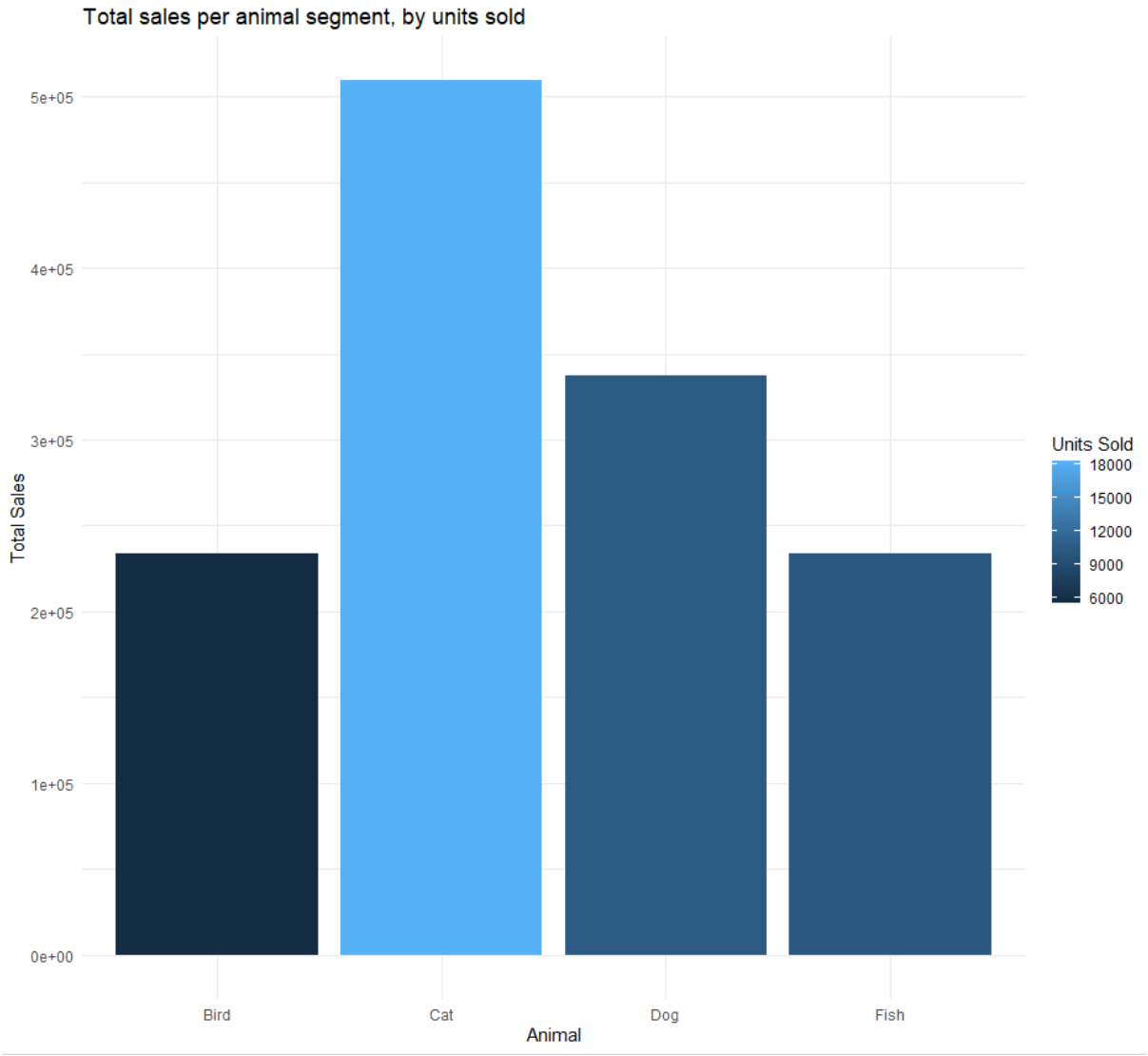


Figure 2.4 – column plot of Total Sales by Animal, gradient with units sold

2.4a) The bird segment has total sales of 234129 and has sold a total of 5459. In relation to 2.1 It follows the original assumptions, that the bird segment has a higher mean sales price, because it sells less units.

2.4b) From the plot and number of units sold it is apparent that the fish segment sells more units at a lower price. While the bird segment sells less units at a higher price. This is consistent with section 2.1 where the bird segment had the highest mean sales, while fish had the lowest.

3. Multiple Linear Regression

Commercial model:

```
> summary(commercial_model)

Call:
lm(formula = sales ~ price + rating + repeat_purchase, data = df_regression)

Residuals:
    Min      1Q  Median      3Q     Max
-321.26  -72.81    0.75   75.98  362.85

Coefficients:
                Estimate Std. Error t value
(Intercept)     -301.8536    14.6652 -20.583
price             37.8928     0.3873  97.847
rating            35.0742     1.8978  18.482
repeat_purchase1   5.2292     6.0068   0.871
                Pr(>|t|)
(Intercept)       <2e-16 ***
price             <2e-16 ***
rating            <2e-16 ***
repeat_purchase1   0.384
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 106.3 on 1325 degrees of freedom
Multiple R-squared:  0.8919,    Adjusted R-squared:  0.8916
F-statistic:  3643 on 3 and 1325 DF,  p-value: < 2.2e-16
```

Positioning model:

```
> summary(positioning_model)

Call:
lm(formula = sales ~ category + animal + size, data = df_regression)

Residuals:
    Min      1Q  Median      3Q     Max
-67.407 -17.635  -2.627  12.401 183.667

Coefficients:
                  Estimate Std. Error t value
(Intercept)       1437.277      3.579  401.54
categoryEquipment  335.332      4.047   82.86
categoryFood       360.463      3.790   95.12
categoryHousing    135.307      3.986   33.94
categoryMedicine   298.279      4.057   73.53
categoryToys       635.016      4.031  157.53
animalCat         -417.138      3.274 -127.41
animalDog         -347.024      3.149 -110.22
animalFish        -737.883      3.262 -226.22
sizemedium        -548.997      2.801 -196.00
sizesmall         -323.374      2.911 -111.10
                  Pr(>|t|)
(Intercept)         <2e-16 ***
categoryEquipment   <2e-16 ***
categoryFood        <2e-16 ***
categoryHousing     <2e-16 ***
categoryMedicine    <2e-16 ***
categoryToys        <2e-16 ***
animalCat           <2e-16 ***
animalDog           <2e-16 ***
animalFish          <2e-16 ***
sizemedium          <2e-16 ***
sizesmall           <2e-16 ***
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 32.24 on 1318 degrees of freedom
Multiple R-squared:  0.9901,    Adjusted R-squared:   0.99
F-statistic: 1.319e+04 on 10 and 1318 DF,  p-value: < 2.2e-16
```

Full model:

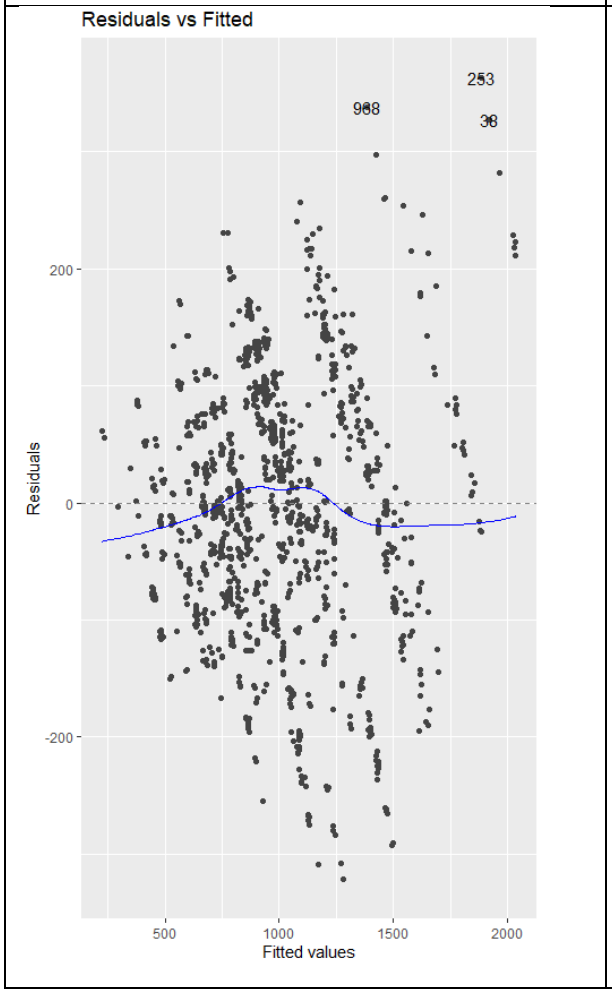
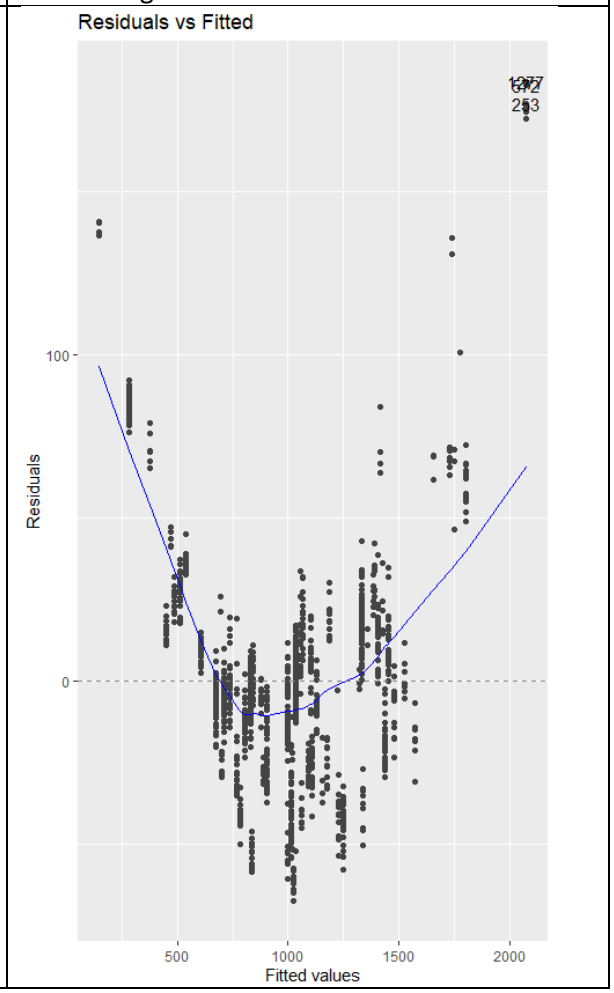
```
> summary(full_model)

Call:
lm(formula = sales ~ ., data = df_regression)

Residuals:
    Min      1Q  Median      3Q     Max
-67.491 -15.055  -2.455  11.518 181.689

Coefficients:
                   Estimate Std. Error t value
(Intercept)       -196.8259   273.8274  -0.719
categoryEquipment  145.4681    32.0523   4.538
categoryFood        56.5458    51.0736   1.107
categoryHousing     21.4785    19.4788   1.103
categoryMedicine   222.3946    13.3087  16.710
categoryToys       216.6206    70.2030   3.086
animalCat          151.3859    95.4195   1.587
animalDog           32.5193    63.7134   0.510
animalFish          21.0331   127.2781   0.165
sizemedium        -169.3840    63.6663  -2.660
sizesmall         -133.2304    31.9785  -4.166
price               37.9632     6.3661   5.963
rating               0.3418     0.6456   0.529
repeat_purchase1     0.1890     1.8065   0.105
                  Pr(>|t|)
(Intercept)        0.47239
categoryEquipment 6.18e-06 ***
categoryFood       0.26843
categoryHousing    0.27037
categoryMedicine   < 2e-16 ***
categoryToys       0.00207 **
animalCat          0.11286
animalDog          0.60986
animalFish         0.86877
sizemedium         0.00790 **
sizesmall         3.30e-05 ***
price             3.17e-09 ***
rating             0.59663
repeat_purchase1   0.91670
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 31.85 on 1315 degrees of freedom
Multiple R-squared:  0.9904,    Adjusted R-squared:  0.9903
F-statistic: 1.04e+04 on 13 and 1315 DF,  p-value: < 2.2e-16
```

Commercial Model	Positioning Model
	

Residual vs. fitted – Commercial model (left), Positioning model (right)

	Commercial model	Positioning model	Full model
Adjusted R_squared	0.8916	0.99	0.9903

3a) from the adjusted R-squares of the different models (see above) the adjusted R-squared ranks the models in the following order 1. Full model, 2. Positioning model, 3. Commercial model. With virtually no increase between the positioning model and the full model it cannot be said that any predictive power from the commercial model adds to the positioning model.

3b) Rating is not a significant variable in the full model. The increase in units sold shown in 2.3 must therefore come from other variables than rating.

3c) Residual vs. fitted tests for linearity and homoscedasticity (constant variance). The residuals shows a curved pattern in the position, indication non-linearity. Thus there are relationships between the variables in the model that cannot be captured by linear models. There are a few notable outliers but not so extreme as to warrant extra check. The linearity assumption can not be said to be satisfied.

3d) The managers claim (b) is incorrect, where the model in fact shows that the opposite is correct. A restated claim should be “What a product earns comes down to Category, Animal and Size. Price and rating matters much less”

4. Advanced Machine Learning: KNN and Random Forest

KNN Confusion Matrix

```
Confusion Matrix and Statistics

          Reference
Prediction Bird Cat Dog Fish
      Bird   31   0   0    0
      Cat     0  90   8    4
      Dog     2   7  51    1
      Fish    0   3   5   62

Overall Statistics

               Accuracy : 0.8864
                 95% CI : (0.8418, 0.922)
    No Information Rate : 0.3788
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.8412

 Mcnemar's Test P-Value : NA

Statistics by Class:

                     Class: Bird Class: Cat Class: Dog Class: Fish
Sensitivity               0.9394     0.9000     0.7969      0.9254
Specificity               1.0000     0.9268     0.9500      0.9594
Pos Pred Value            1.0000     0.8824     0.8361      0.8857
Neg Pred Value            0.9914     0.9383     0.9360      0.9742
Prevalence                0.1250     0.3788     0.2424      0.2538
Detection Rate            0.1174     0.3409     0.1932      0.2348
Detection Prevalence      0.1174     0.3864     0.2311      0.2652
Balanced Accuracy         0.9697     0.9134     0.8734      0.9424
```

Random Forest confusion Matrix

```
Confusion Matrix and Statistics

          Reference
Prediction Bird Cat Dog Fish
      Bird   33   0   0    0
      Cat     0  91   3    2
      Dog     0   7  60    0
      Fish    0   2   1   65

Overall Statistics

               Accuracy : 0.9432
                 95% CI : (0.908, 0.9679)
    No Information Rate : 0.3788
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.9211

 Mcnemar's Test P-Value : NA

Statistics by Class:

                     Class: Bird Class: Cat Class: Dog Class: Fish
Sensitivity                1.000      0.9100      0.9375       0.9701
Specificity                1.000      0.9695      0.9650       0.9848
Pos Pred Value             1.000      0.9479      0.8955       0.9559
Neg Pred Value             1.000      0.9464      0.9797       0.9898
Prevalence                 0.125      0.3788      0.2424       0.2538
Detection Rate             0.125      0.3447      0.2273       0.2462
Detection Prevalence       0.125      0.3636      0.2538       0.2576
Balanced Accuracy          1.000      0.9398      0.9512       0.9775
```



Variable importance plot of Random Forest model

Q3a. *Report the accuracy of both models and compare it to the No Information Rate, which is printed below the accuracy in the console output. Are the commercial variables sufficient to recover the animal segment?*

The No Information Rate is 0.3788 for both models, as they are trained and tested on the same sets. The no information rate is the percentage of observation correctively predicted if only picking the majority class. With an accuracy of 0.8864 and 0.9432 for knn and Random forest respectively well above the no information rate, it can be concluded that the commercial variables is sufficient to recover the animal segment. This is further corroborated by a p-value for [Accuracy > NIR] well under 0.05 (5% significance-level) for both models.

Q3b. *Inspect the class-wise statistics of each model. Which animal segment does the KNN model identify worst, and does the random forest have the same weakness? Briefly explain what may cause it.*

Seen from the sensitivity (true-positive rate) of the classes in the knn model (see below), the knn model identifies the dog-segment the worst. This is not the case in the random forest model. One explanation is that the random forest uses many decorrelated trees to find the optimal model.

Sensitivity \ animal	Bird	Cat	Dog	Fish
KNN	0.9394	0.9000	0.7969	0.9254
Random Forest	1.00	0.9100	0.9375	0.9701

Q3c. Why is centering and scaling of the predictors essential for KNN, while the random forest does not require it?

Centering and Scaling is necessary for the knn model as it uses relative distance between variables to make predictions. Without centering and scaling variables with greater relative distances between observations would influence the model unfairly. The random forest model on the other hand uses decision rules to predict and is not subject to influence from distance

Q3d. Read the variable importance chart. Which predictor contributes least, and is this consistent with what you found in Section 3?

The variable importance plot shows that Rating is the least important variable for prediction of the Animal Segment. While the models predict for different response variables it is consistent with section 3. And how rating has little predictive power.

5. Management Advice

For the section below it is assumed that it is 5-7 sentences in total rather than for each section, as

Item	Result
Claim(a)	Original claim is **partially false** – While average per unit sales are higher for Birds, it does not generate the most sales overall. In fact the segment generates the least revenue overall. Where the marketing budget should move depends on multiple factors. However due to the high price median from section 2.2 and low unit sales from 2.4 boosting sales in Birds would help Nordpaw Pet Supplies.
Claim(b)	Original claim is **false** - The analysis has shown that rating and price and rating are not important for sales. Rather the category, animal segment and size decide the sales within Nordpaw Pet Supplies.
Segment classification for onboarding	The onboarding team can **rely on automated segment classifiers** for deciding which segment a sale belongs to. This is seen through a performance well over the “no information rate” both for the KNN and Random forest model, in section 4.