AVONET EDA

Collin Van Allen

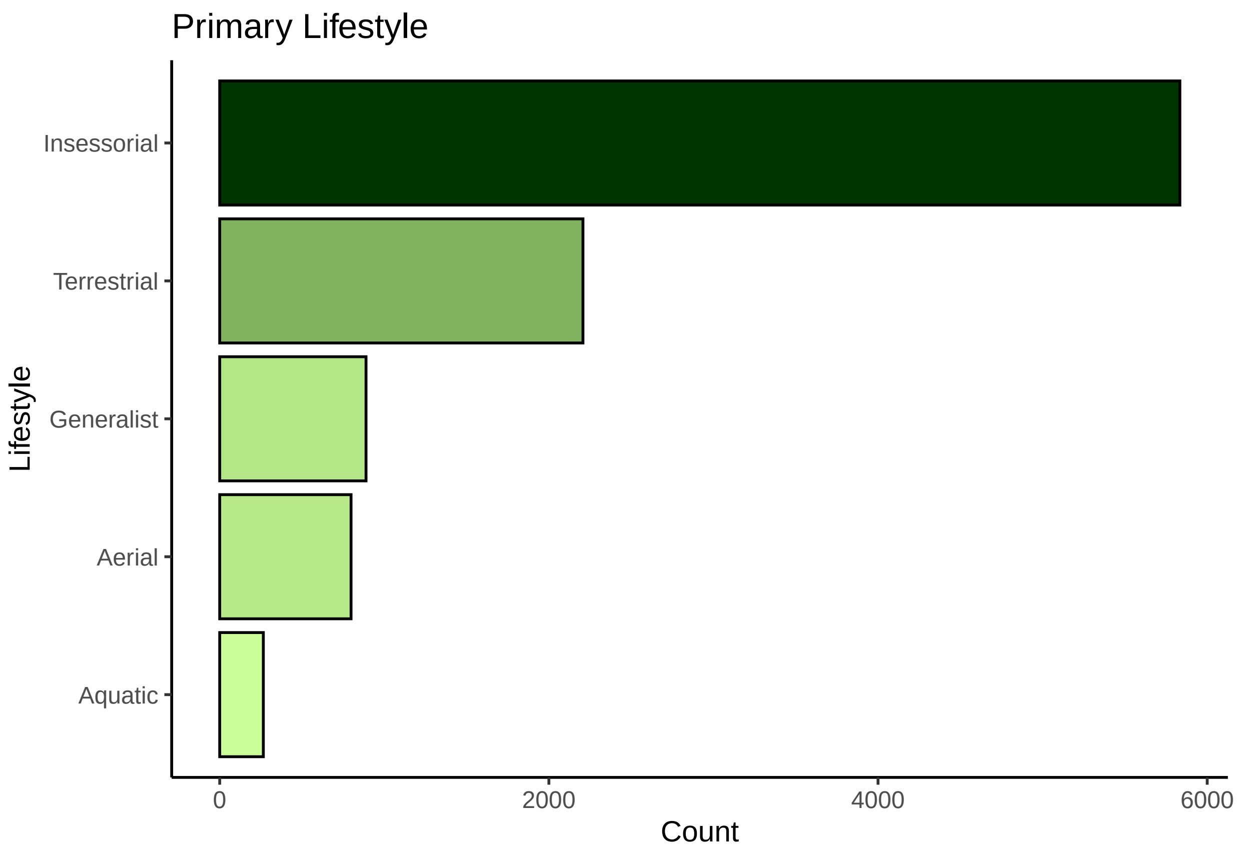
2022-12-17

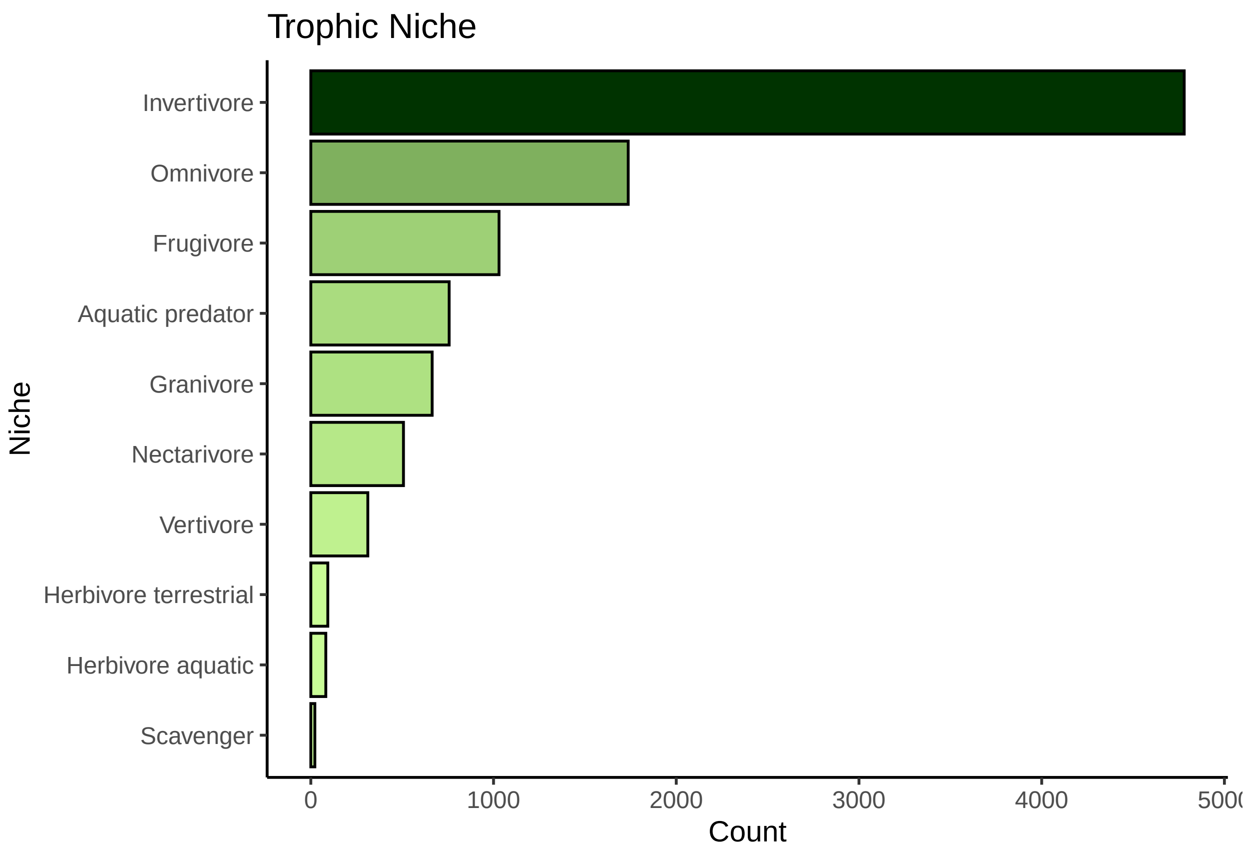
AVONET

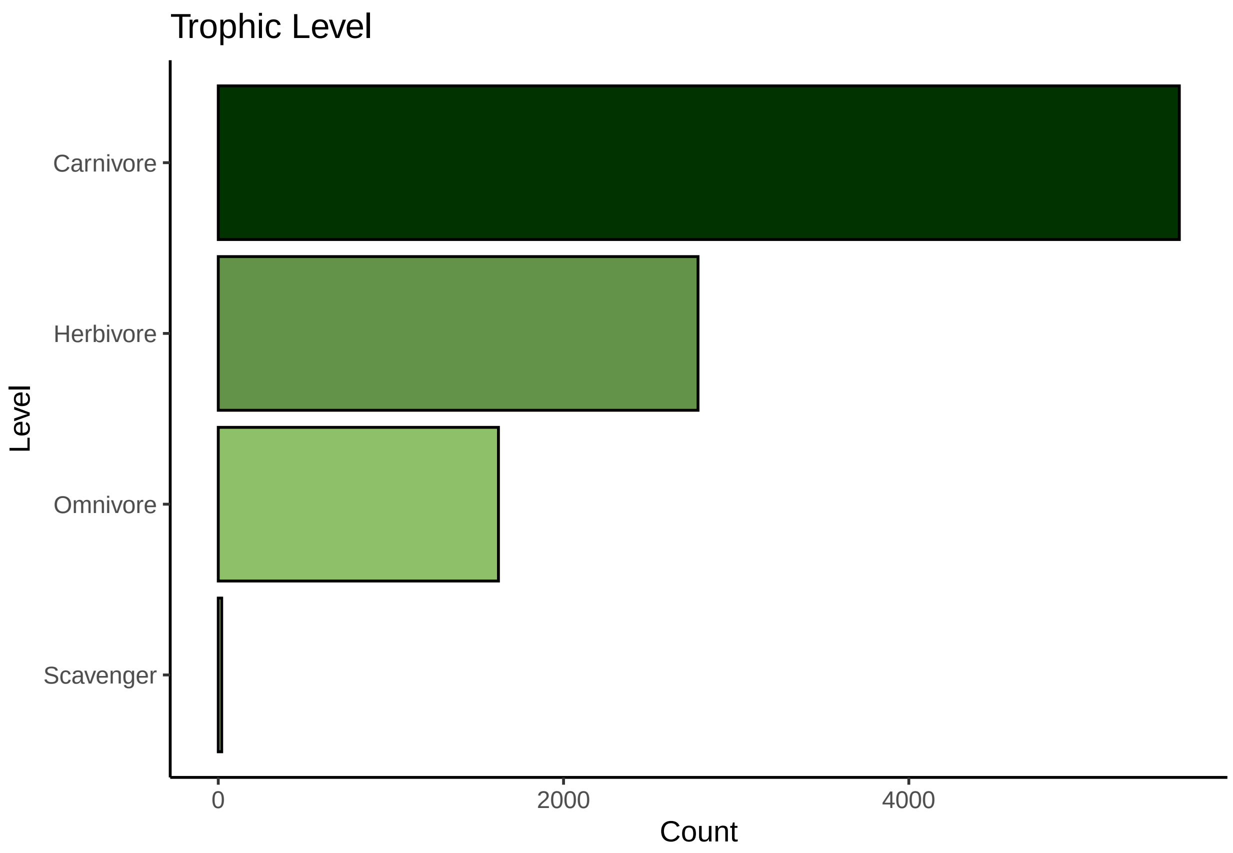
AVONET is a data set that was created to compile data on birds from multiple sources. The data set encompasses over 11,000 different species with 11 morphological traits recorded for each species. The data set contains average measurements for all the species listed. This project is only including one of the sheets located in the supplementary data set 1 and only contains roughly 10,000 of the birds from the full set. Information will also be slightly different from the main, compiled, data set since it is only from one source. An article for the data set can be found here: https://onlinelibrary.wiley.com/doi/full/10.1111/ele.13898. Downloads for the data set can be found in the article or at this link: https://figshare.com/s/b990722d72a26b5bfead.

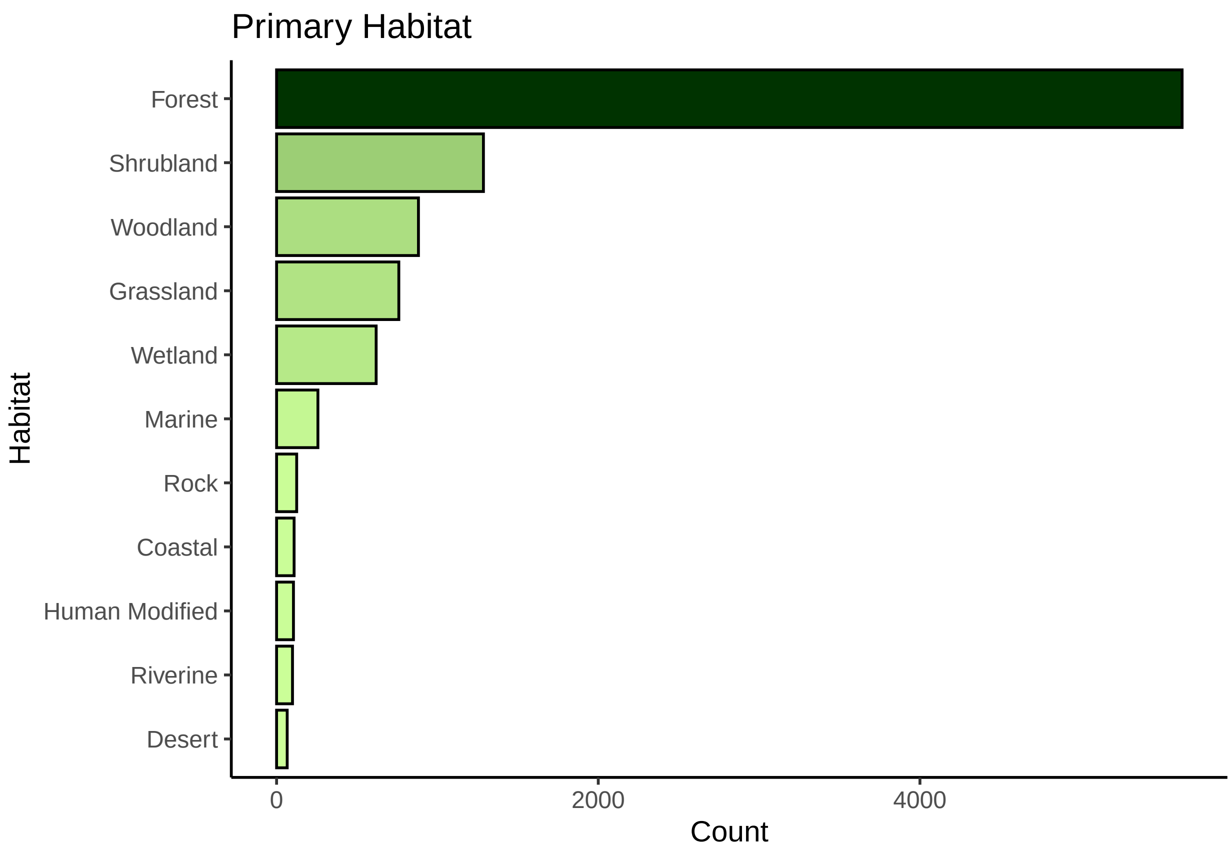
Basic Overview on Categorical Data

The main purpose of these plots on categorical data are to gauge the data set for any factors that may be of interest in the get go. These graphs also serve to determine which categorical features should be used in future graphs and tests.


Primary Lifestyle
Lifestyle
Insessorial
Terrestrial
Generalist
Aerial
Aquatic
0
2000
4000
6000
Count


Trophic Niche
Niche
Invertivore
Omnivore
Frugivore
Aquatic predator
Granivore
Nectarivore
Vertivore
Herbivore terrestrial
Herbivore aquatic
Scavenger
0
1000
2000
3000
4000
500(
Count

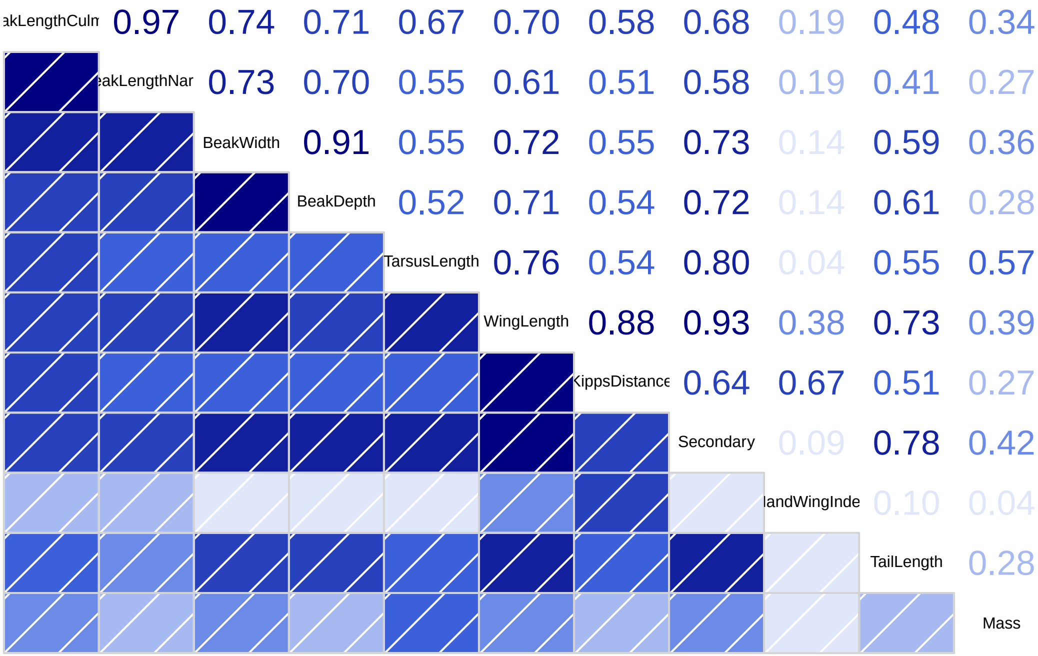

Trophic Level
Carnivore
Herbivore
Omnivore
Scavenger
Level
0
2000
4000
Count



Overall the categorical variables provide an interesting picture for what the collection of bird species looks like. Of these, trophic niche is an interesting one to try and use for predicting certain members of a given taxa. It may pose issues if used to classify or cluster though since it has so many levels in it. Similar story for the habitats of the birds. As a whole they will all be useful in classifying other traits or predicting since they all provide important information about a given bird.

Correlation Plot

Like the categorical graphs this graph serves to find any features of interest to focus on in future graphs and tests.



The correlation plot, like the bar graphs, provide an interesting view into the data set. Of interest I think the measurements that cover the size of the birds beak and wings will provide more information than the other measures. While the other measurements may be useful they either seems negligible in my mind or are used in other measurements, such is the case for Kipp's Distance.

With most of the exploratory data analysis out of the way, there are questions and hypotheses that I want to form with the data.

Questions:

1. Using kNN; Is it possible to predict the trophic niche of a bird?
2. Using a decision tree; Can all the important features (measurements and categorical types) predict the order or family of the bird?
    - Possible follow up: Can this be used to find how closely related some birds are, or the degrees of separation on a phylogenetic tree?
3. Using PCA; Is there a way to find the best measurement features for a bird in a given trophic level/niche, habitat, or primary lifestyle?

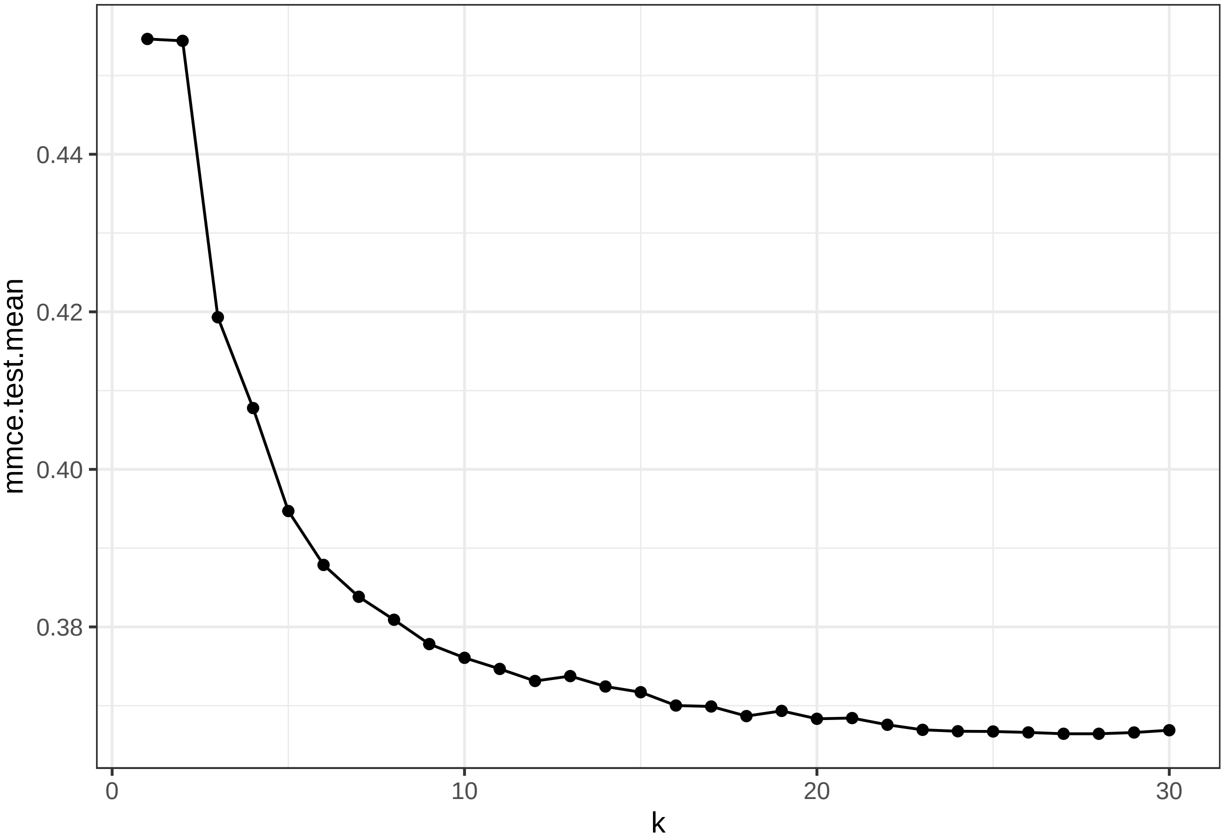
---

kNN Approach

With this I want to find out if it's possible to classify birds by trophic niche using only their beak size. Although not an entirely accurate way to measure since many birds share similar beak measurements but different diets. It is still interesting nonetheless to see if beak size provides a similar measurement to what multiple measurements of a given bird could provide.

Hypothesis: Using mass, beak measurements, wing size, and habitat the trophic niche of a bird can be more accurately predicted than using just the size of the beak and primary lifestyle. Using purely kNN with repeated cross fold validation I created two models. The first model only contains measurements of the bird beak size and trophic niche for classification. The model uses tuned k value which is around 16 and goes through 20 reps of 5 fold validation. The tuning of k was done with 10 folds and 10 reps.

```
## Relative confusion matrix (normalized by row/column):
##                         predicted
## true                    Aquatic predator Frugivore   Granivore  
##   Aquatic predator      5e-01/6e-01      3e-02/5e-02 6e-04/7e-04
##   Frugivore             2e-02/3e-02      2e-01/4e-01 4e-02/7e-02
##   Granivore             9e-04/9e-04      4e-02/5e-02 6e-01/6e-01
##   Herbivore aquatic     1e-01/2e-02      7e-02/1e-02 0e+00/0e+00
##   Herbivore terrestrial 5e-02/6e-03      1e-01/2e-02 6e-02/9e-03
##   Invertivore           2e-02/2e-01      2e-02/2e-01 7e-03/5e-02
##   Nectarivore           2e-02/2e-02      1e-02/1e-02 1e-02/9e-03
##   Omnivore              6e-02/2e-01      6e-02/2e-01 8e-02/2e-01
##   Scavenger             7e-02/2e-03      3e-01/1e-02 0e+00/0e+00
##   Vertivore             6e-03/3e-03      5e-02/4e-02 2e-02/1e-02
##   -err.-                       0.41             0.60        0.38
##                         predicted
## true                    Herbivore aquatic Herbivore terrestrial Invertivore
##   Aquatic predator      2e-02/2e-01       6e-03/2e-01           3e-01/3e-02
##   Frugivore             2e-03/2e-02       4e-03/1e-01           6e-01/1e-01
##   Granivore             0e+00/0e+00       2e-04/5e-03           2e-01/2e-02
##   Herbivore aquatic     4e-01/4e-01       3e-02/8e-02           2e-01/2e-03
##   Herbivore terrestrial 7e-02/9e-02       1e-01/3e-01           3e-01/4e-03
##   Invertivore           2e-03/1e-01       7e-04/1e-01           9e-01/7e-01
##   Nectarivore           0e+00/0e+00       0e+00/0e+00           1e-01/9e-03
##   Omnivore              8e-03/2e-01       4e-03/2e-01           5e-01/1e-01
##   Scavenger             0e+00/0e+00       0e+00/0e+00           5e-02/2e-04
##   Vertivore             0e+00/0e+00       3e-03/3e-02           2e-01/8e-03
##   -err.-                       0.59              0.70                  0.31
##                         predicted
## true                    Nectarivore Omnivore    Scavenger   Vertivore  
##   Aquatic predator      2e-02/3e-02 1e-01/7e-02 1e-03/8e-02 2e-02/5e-02
##   Frugivore             1e-04/2e-04 1e-01/1e-01 5e-04/4e-02 4e-02/1e-01
##   Granivore             1e-02/1e-02 2e-01/1e-01 0e+00/0e+00 1e-03/2e-03
##   Herbivore aquatic     0e+00/0e+00 2e-01/1e-02 5e-03/3e-02 0e+00/0e+00
##   Herbivore terrestrial 5e-04/1e-04 3e-01/2e-02 0e+00/0e+00 7e-02/2e-02
##   Invertivore           7e-03/7e-02 5e-02/2e-01 1e-04/5e-02 1e-02/1e-01
##   Nectarivore           8e-01/8e-01 9e-02/4e-02 0e+00/0e+00 1e-03/2e-03
##   Omnivore              2e-02/8e-02 3e-01/4e-01 2e-03/2e-01 4e-02/2e-01
##   Scavenger             0e+00/0e+00 2e-01/4e-03 3e-01/6e-01 5e-02/3e-03
##   Vertivore             1e-03/7e-04 1e-01/3e-02 1e-03/3e-02 6e-01/5e-01
##   -err.-                        0.19        0.61        0.45        0.47
##                        predicted
## true                    -err.-
##   Aquatic predator      0.49
##   Frugivore             0.82
##   Granivore             0.43
##   Herbivore aquatic     0.62
##   Herbivore terrestrial 0.90
##   Invertivore           0.11
##   Nectarivore           0.24
##   Omnivore              0.73
##   Scavenger             0.68
##   Vertivore             0.38
##   -err.-                0.37
##
##
## Absolute confusion matrix:
##                        predicted
## true                    Aquatic predator Frugivore Granivore Herbivore aquatic
##   Aquatic predator                  7678       488         9               301
##   Frugivore                          389      3669       821                37
##   Granivore                           12       494      7542                 0
##   Herbivore aquatic                  235       111         0               625
##   Herbivore terrestrial               84       195       108               137
##   Invertivore                       2036      1678       659               153
##   Nectarivore                        238        99       107                 0
##   Omnivore                          2186      1929      2713               277
##   Scavenger                           29       133         0                 0
##   Vertivore                           37       323       138                 0
##   -err.-                            5246      5450      4555               905
##                        predicted
## true                    Herbivore terrestrial Invertivore Nectarivore Omnivore
##   Aquatic predator                         96        4140         276     1787
##   Frugivore                                73       11965           2     2899
##   Granivore                                 3        2593         127     2492
##   Herbivore aquatic                        47         296           0      318
##   Herbivore terrestrial                   188         530           1      487
##   Invertivore                              65       84682         630     4727
##   Nectarivore                               0        1082        7718      884
##   Omnivore                                127       16008         813     9390
##   Scavenger                                 0          22           0       91
##   Vertivore                                19        1010           7      821
##   -err.-                                  430       37646        1856    14506
##                        predicted
## true                    Scavenger Vertivore -err.-
##   Aquatic predator             20       345   7462
##   Frugivore                    11       734  16931
##   Granivore                     0        17   5738
##   Herbivore aquatic             8         0   1015
##   Herbivore terrestrial         0       130   1672
##   Invertivore                  13       957  10918
##   Nectarivore                   0        12   2422
##   Omnivore                     55      1242  25350
##   Scavenger                   142        23    298
##   Vertivore                    7      3878   2362
##   -err.-                     114      3460  74168
```

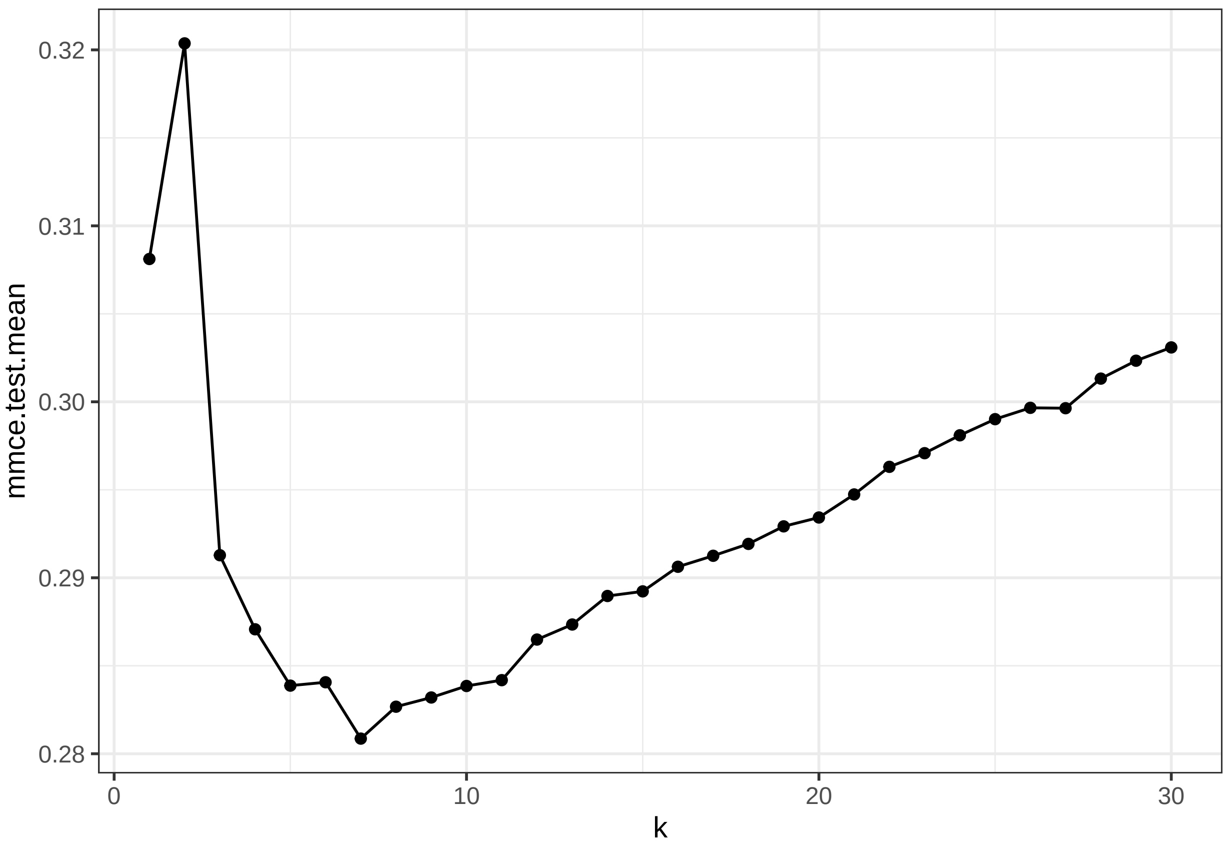


Overall the model did fairly well at classifying the trophic niche of the birds, but it's far from perfect and could be improved. As the model stands now multiple niches are incorrectly identified at rates above 50%.

In the second model I attempted to improve the model by including the wing length (length from carpal joint to wingtip), secondary (length from carpal joint to outermost secondary feather), and habitat. The idea is that when beak measurements are not enough, a higher accuracy should be achievable if the model can learn the size of the wing and habitat in which the bird lives.

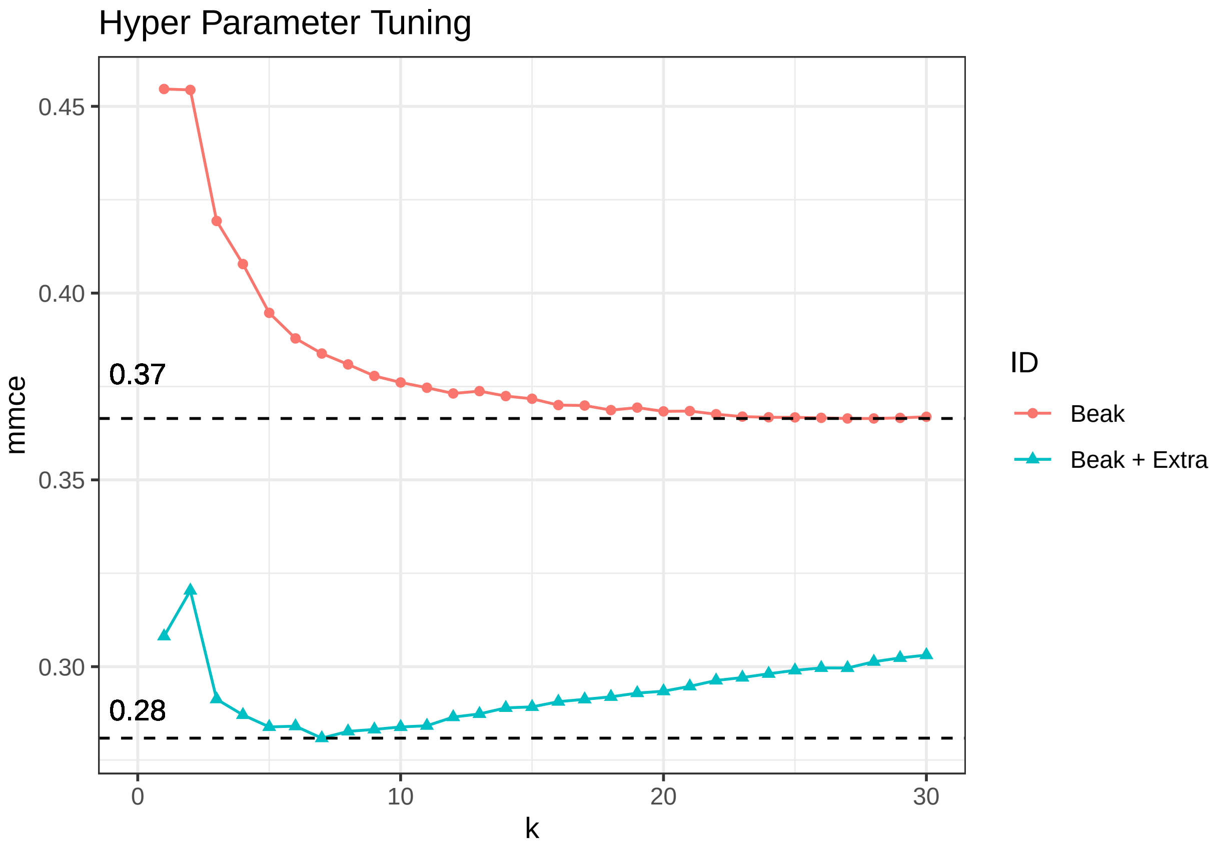
In this model almost all the parameters and cross fold validation are the same bar the k for this model. After tuning the model it was found that the optimal value for k is around 8.

```
## Relative confusion matrix (normalized by row/column):
##                         predicted
## true                     Aquatic predator Frugivore   Granivore
##   Aquatic predator       8e-01/8e-01      6e-03/6e-03 0e+00/0e+00
##   Frugivore              9e-04/1e-03      5e-01/6e-01 3e-02/4e-02
##   Granivore              3e-03/3e-03      5e-02/4e-02 6e-01/7e-01
##   Herbivore aquatic      1e-01/2e-02      0e+00/0e+00 6e-04/8e-05
##   Herbivore terrestrial 3e-02/4e-03      6e-02/6e-03 2e-02/3e-03
##   Invertivore            9e-03/6e-02      3e-02/2e-01 9e-03/7e-02
##   Nectarivore            0e+00/0e+00      2e-02/1e-02 1e-02/8e-03
##   Omnivore               4e-02/9e-02      1e-01/2e-01 7e-02/2e-01
##   Scavenger              1e-01/3e-03      5e-03/1e-04 0e+00/0e+00
##   Vertivore              7e-03/3e-03      3e-02/1e-02 2e-04/8e-05
##   -err.-                       0.18             0.45        0.32
##                        predicted
## true                    Herbivore aquatic Herbivore terrestrial Invertivore
##   Aquatic predator      3e-02/2e-01       9e-04/1e-02           8e-02/1e-02
##   Frugivore             2e-04/2e-03       1e-03/2e-02           3e-01/6e-02
##   Granivore             8e-04/5e-03       7e-04/9e-03           1e-01/1e-02
##   Herbivore aquatic     7e-01/5e-01       2e-02/3e-02           2e-02/4e-04
##   Herbivore terrestrial 1e-02/1e-02       3e-01/5e-01           1e-01/2e-03
##   Invertivore           7e-04/3e-02       1e-03/1e-01           9e-01/8e-01
##   Nectarivore           0e+00/0e+00       0e+00/0e+00           1e-01/1e-02
##   Omnivore              1e-02/2e-01       7e-03/2e-01           3e-01/1e-01
##   Scavenger             1e-02/3e-03       0e+00/0e+00           0e+00/0e+00
##   Vertivore             0e+00/0e+00       1e-02/9e-02           1e-01/8e-03
##   -err.-                       0.46              0.51                  0.22
##                        predicted
## true                    Nectarivore Omnivore    Scavenger   Vertivore
##   Aquatic predator      2e-03/3e-03 4e-02/2e-02 2e-03/8e-02 2e-02/4e-02
##   Frugivore             3e-03/7e-03 2e-01/1e-01 0e+00/0e+00 2e-02/4e-02
##   Granivore             0e+00/0e+00 2e-01/1e-01 0e+00/0e+00 7e-04/1e-03
##   Herbivore aquatic     0e+00/0e+00 1e-01/6e-03 7e-03/4e-02 2e-02/4e-03
##   Herbivore terrestrial 0e+00/0e+00 4e-01/3e-02 0e+00/0e+00 1e-01/3e-02
##   Invertivore           4e-03/4e-02 5e-02/2e-01 0e+00/0e+00 1e-02/1e-01
##   Nectarivore           8e-01/9e-01 8e-02/3e-02 0e+00/0e+00 0e+00/0e+00
##   Omnivore              2e-02/8e-02 4e-01/5e-01 2e-03/2e-01 3e-02/1e-01
##   Scavenger             0e+00/0e+00 1e-02/2e-04 4e-01/6e-01 5e-01/3e-02
##   Vertivore             0e+00/0e+00 5e-02/1e-02 4e-03/8e-02 8e-01/6e-01
##   -err.-                       0.13        0.51        0.43        0.39
##                        predicted
## true                    -err.-
##   Aquatic predator      0.18
##   Frugivore             0.53
##   Granivore             0.37
##   Herbivore aquatic     0.30
##   Herbivore terrestrial 0.73
##   Invertivore           0.12
##   Nectarivore           0.24
##   Omnivore              0.61
##   Scavenger             0.60
##   Vertivore             0.25
##   -err.-                0.29
##
##
## Absolute confusion matrix:
##                        predicted
## true                    Aquatic predator Frugivore Granivore Herbivore aquatic
##   Aquatic predator                 12235        97         0               509
##   Frugivore                           18      9624       521                 4
##   Granivore                           46       645      8320                11
##   Herbivore aquatic                  228         0         1              1142
##   Herbivore terrestrial               58       107        43                24
##   Invertivore                        867      3003       833                64
##   Nectarivore                          0       182        99                 0
##   Omnivore                          1331      3558      2499               365
##   Scavenger                           44         2         0                 6
##   Vertivore                           42       167         1                 0
##   -err.-                            2634      7761      3997               983
##                         predicted
## true                    Herbivore terrestrial Invertivore Nectarivore Omnivore
##   Aquatic predator                         14        1122          24      597
##   Frugivore                                20        6538          62     3472
##   Granivore                                 9        1566           0     2574
##   Herbivore aquatic                        31          38           0      161
##   Herbivore terrestrial                   503         183           0      706
##   Invertivore                              98       83891         382     5050
##   Nectarivore                               0        1343        7703      793
##   Omnivore                                253       11351         731    13365
##   Scavenger                                 0           0           0        5
##   Vertivore                                91         905           0      314
##   -err.-                                  516       23046        1199    13672
##                         predicted
## true                    Scavenger Vertivore -err.-
##   Aquatic predator             24       318   2705
##   Frugivore                     0       321  10956
##   Granivore                     0         9   4860
##   Herbivore aquatic            11        28    498
##   Herbivore terrestrial         0       236   1357
##   Invertivore                   0       932  11229
##   Nectarivore                   0         0   2417
##   Omnivore                     72       995  21155
##   Scavenger                   177       206    263
##   Vertivore                    25      4695   1545
##   -err.-                      132      3045  56985
```



The results of the model with added features improved the classification power of the model overall. There are less classifications that are above 50% and the over error of the model is much lower than the first model.

kNN Results



By taking both graphs of the tuned hyperparamter we can see that the second model always out performs the first model and is overall much better at predicting the niche of the birds. In summary the hypothesis that I proposed turned out to be correct and the model with more measurements than the ones that measure the beak is better.

Decision Tree Model

For this model I wanted to try and classify the taxonomic family of given birds. I wanted to try a decision tree for this model because can handle multiple classes, and in a sense, it resembles the structure of a phylogenetic tree. With the amount of measurements and birds, it may struggle to classify birds.

Hypothesis: Using purely measurements on the bird such as beak size, wing size, and tarsus length, the rate of misclassification will be lower for a model containing all measurements and the habitat and lifestyle of the bird versus just the base measurements of the bird.

```
## Tune result:
## Op. pars: minsplit=14; minbucket=3; cp=0.0113; maxdepth=6
## mmce.test.mean=0.7929561
```

```
## Resample Result
## Task: bird.family
## Learner: classif.rpart.tuned
## Aggr perf: mmce.test.mean=0.7939654
## Runtime: 434.831
```

Given the nature of decision trees and the large number of families that it needs to account for, there is less of a visual aspect to what the tests are running. With that being said, it still produces good information. We can see that just using the measurements of a bird are not enough to classify the family. It is actually quite poor at classifying the family without knowing the habitat or diet of the bird.

```
## Tune result:
## Op. pars: minsplit=19; minbucket=10; cp=0.0113; maxdepth=7
## mmce.test.mean=0.8096704
```

```
## Resample Result
## Task: bird.familyex
## Learner: classif.rpart.tuned
## Aggr perf: mmce.test.mean=0.7925644
## Runtime: 451.048
```

The second model with more variables did just barely worse than the original model. Overall both models did a very poor job at classifying the family of the bird. My hypothesis for this set of models was wrong as the base measurements did better than the extended model. If I had to choose a better model I would probably choose something like a support vector machine. With that being said, due to the size of the data set and variables I could foresee problems with running the model since it is so intensive with its resources.

Using PCA

For my final question I wanted to look at the the measurements for the birds and how they relate to some of the categorical data. Ideally I would have done this at the start to create better models in the kNN and decision tree models, but I wanted to approach those blindly. With these PCA models I can get a better understanding of why certain models might not have worked as well and have better information for future tests and models.

Hypothesis: When looking at the PCA models, the first 2 components of each model will do better at explaining the variance in both trophic niche and taxonomic family of the birds than the kNN and decision tree did. While the PCA models and classification algorithms are not easily comparable, the PCA models will provide a better picture of the measurements that go in to predicting the trophic niche and family. In the first model I wanted to put all the measurements into the PCA model and see how they contribute to the trophic niche of the birds. When I looked at these in the kNN model it was found that all the measurements, with the inclusion of habitat, produced mmce values below 30%.

```
## [1] "Here are the eigenvectors for the 6 PCS"
```

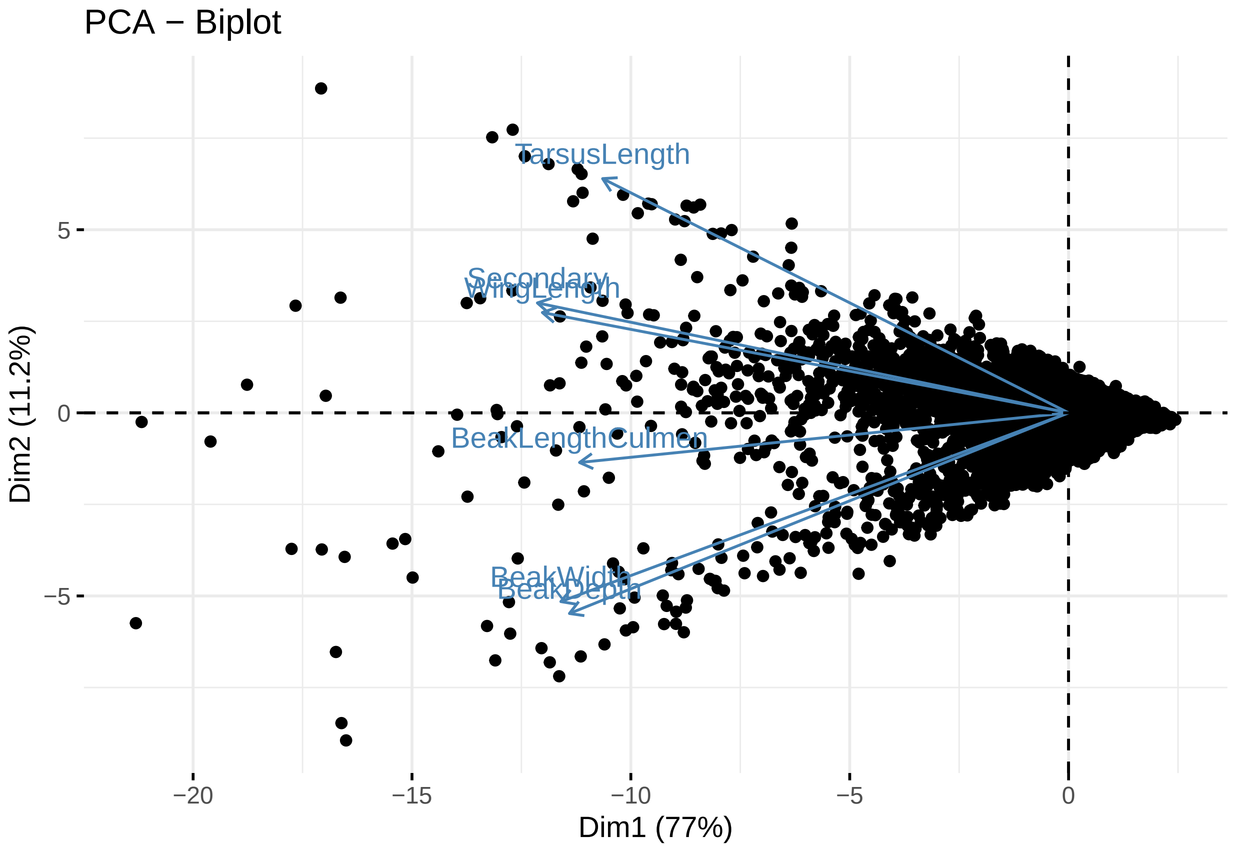
```
##                          PC1        PC2         PC3        PC4           PC5
## BeakLengthCulmen -0.3963710 -0.1260253 -0.80044898 -0.4050239  0.0683378679
## BeakWidth        -0.4115066 -0.4788716  0.09415683  0.2333676 -0.7330278890
## BeakDepth        -0.4045657 -0.5091471  0.17971433  0.2788351  0.6753339376
## TarsusLength     -0.3776767  0.5941787 -0.25841095  0.6367019 -0.0002182158
## WingLength       -0.4264816  0.2545380  0.33419370 -0.5250433 -0.0294641138
## Secondary        -0.4305261  0.2788026  0.37370715 -0.1506202  0.0324952416
##                           PC6
## BeakLengthCulmen -0.13256968
## BeakWidth         0.02623893
## BeakDepth         0.10475515
## TarsusLength      0.17929402
## WingLength        0.60423492
## Secondary        -0.75730904
```

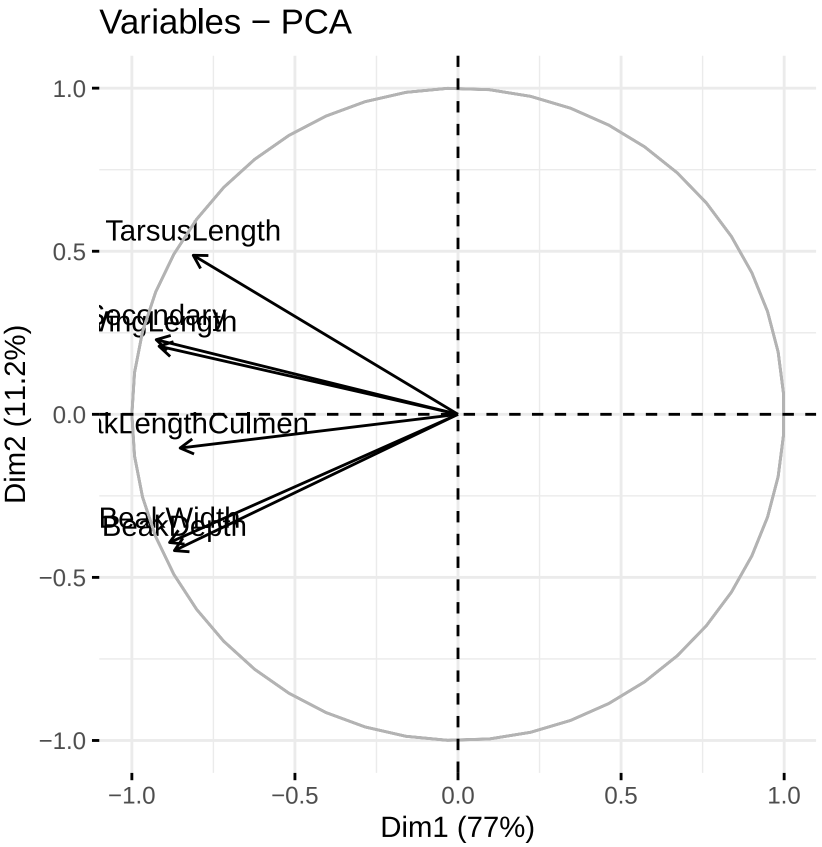
```
## [1] "And here are the square roots of the eigenvalues"
```

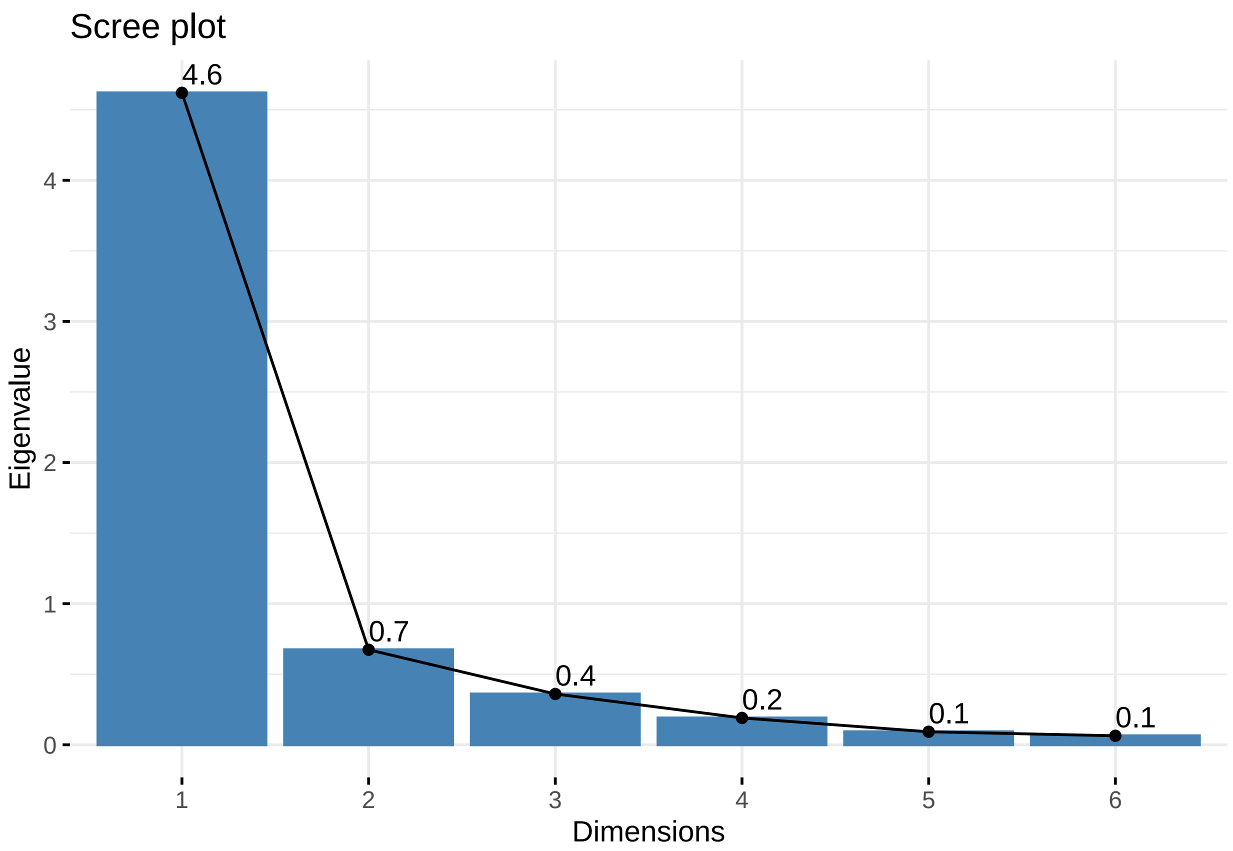
```
## [1] 2.1493670 0.8207036 0.6000794 0.4362145 0.3039451 0.2527968
```

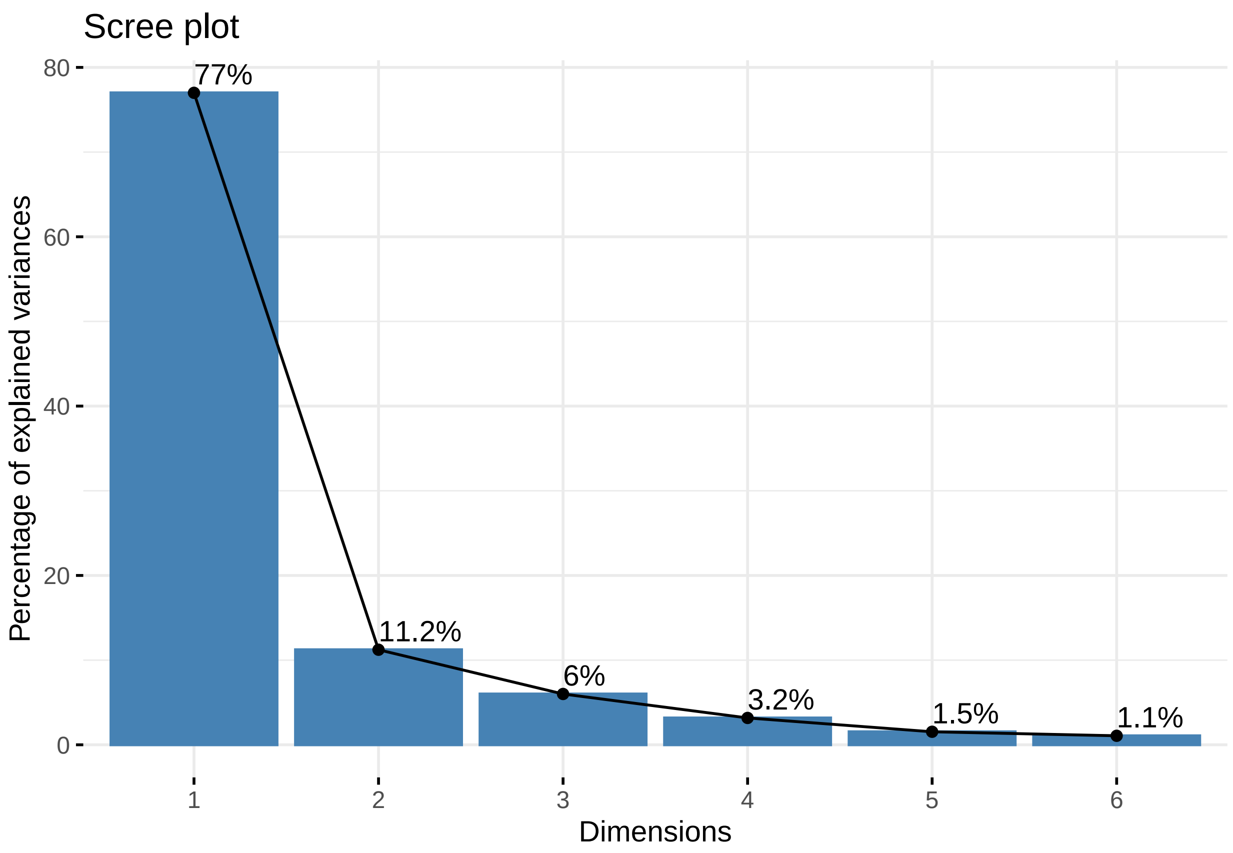
```
## [1] "These are the loadings"
```

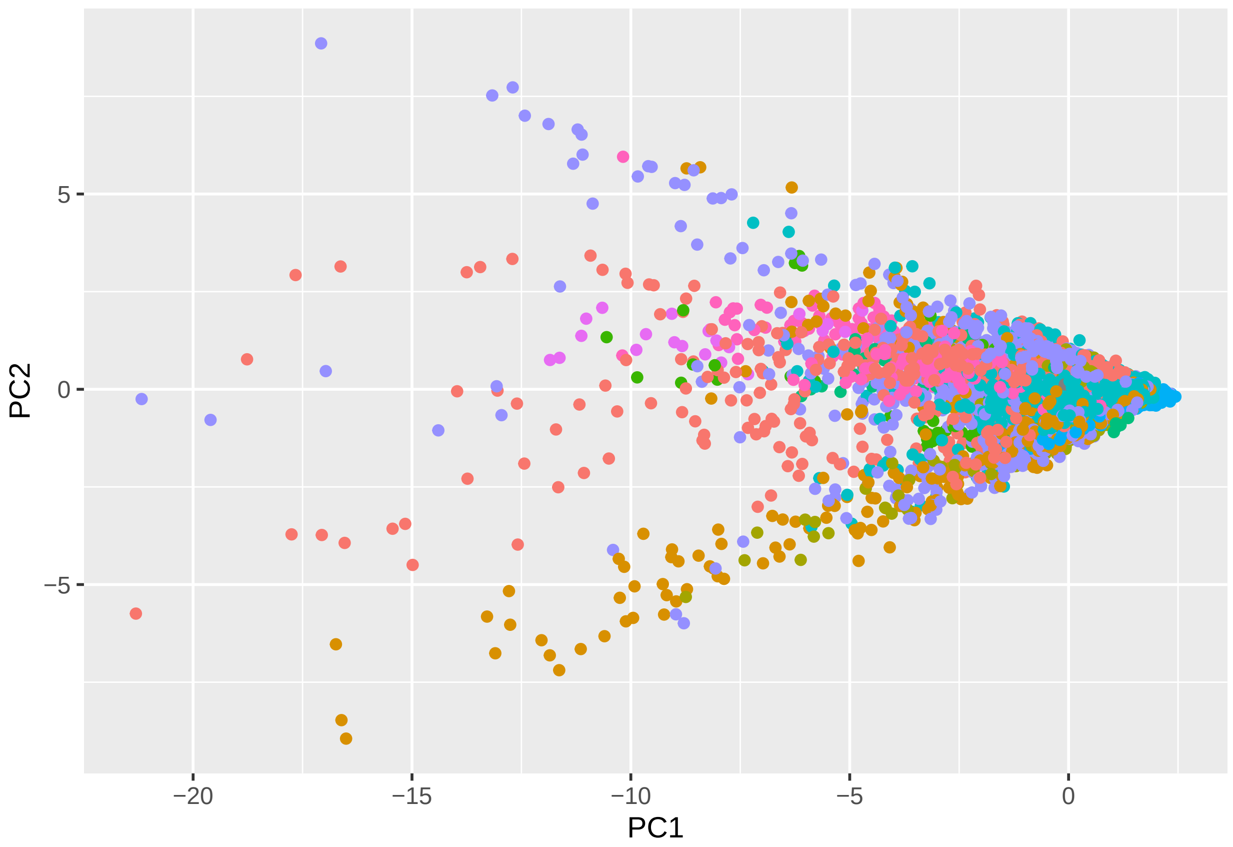
```
## # A tibble: 6 x 6
##     ...1   ...2   ...3   ...4       ...5     ...6
##    <dbl>  <dbl>  <dbl>  <dbl>      <dbl>    <dbl>
## 1 -0.852 -0.103 -0.480  -0.177   0.0208    -0.0335
## 2 -0.884 -0.393  0.0565  0.102  -0.223      0.00663
## 3 -0.870 -0.418  0.108   0.122   0.205      0.0265
## 4 -0.812  0.488 -0.155   0.278  -0.0000663  0.0453
## 5 -0.917  0.209  0.201  -0.229  -0.00896    0.153
## 6 -0.925  0.229  0.224  -0.0657  0.00988   -0.191
```




Variables − PCA
TarsusLength
Secondary
WingLength
kLengthCulmen
BeakWidth
BeakDepth
1.0
0.5
0.0
−0.5
−1.0
Dim2 (11.2%)
−1.0
−0.5
0.0
0.5
1.0
Dim1 (77%)


Scree plot
4.6
0.7
0.4
0.2
0.1
0.1
Eigenvalue
0
1
2
3
4
1
2
3
4
5
6
Dimensions


Scree plot
Percentage of explained variances
80
60
40
20
0
77%
11.2%
6%
3.2%
1.5%
1.1%
1
2
3
4
5
6
Dimensions

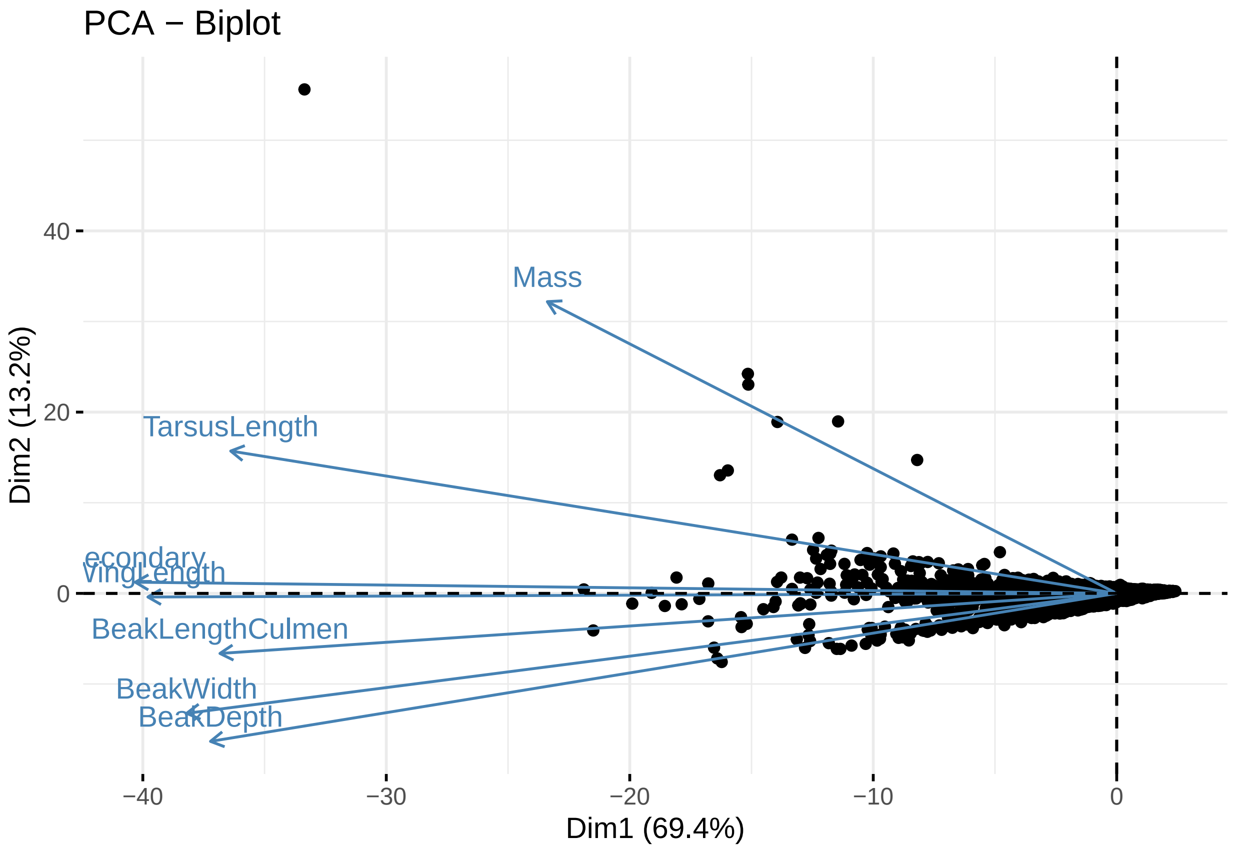


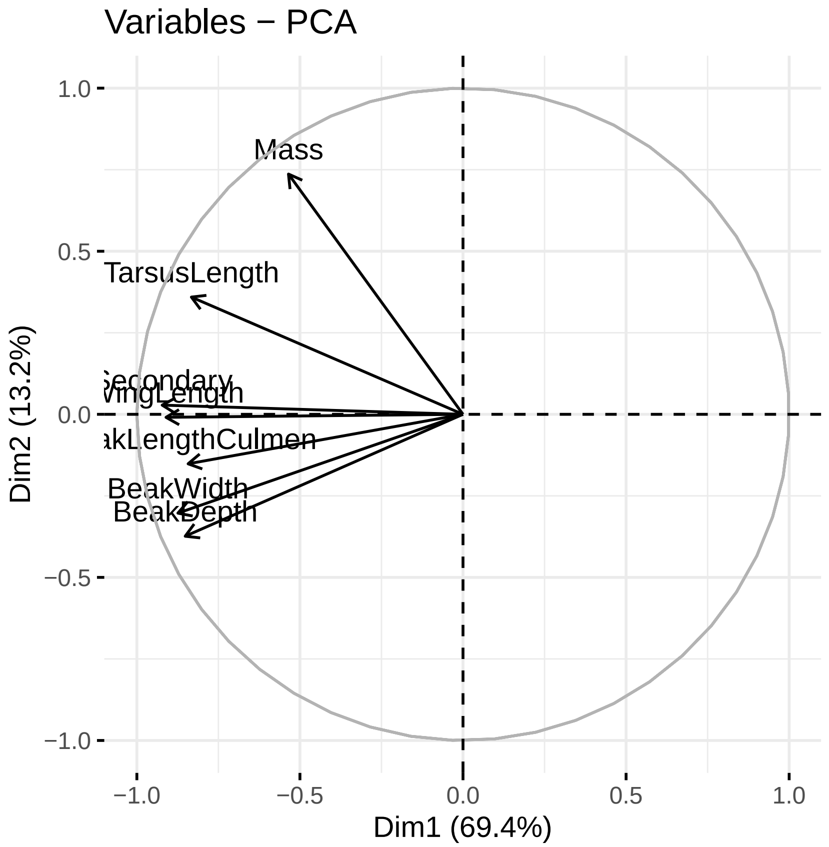
When looking at the first PCA focused on trophic level we can see that almost 90% of the variance in trophic level can be explained by the first two components. This is surprising to see since the mmce values in the kNN model suggest that these variables do not do as well when classifying. It is also interesting to see that the beak measurements provide more of a negative effect towards the classification of trophic niche than the wing and tarsus measurements.

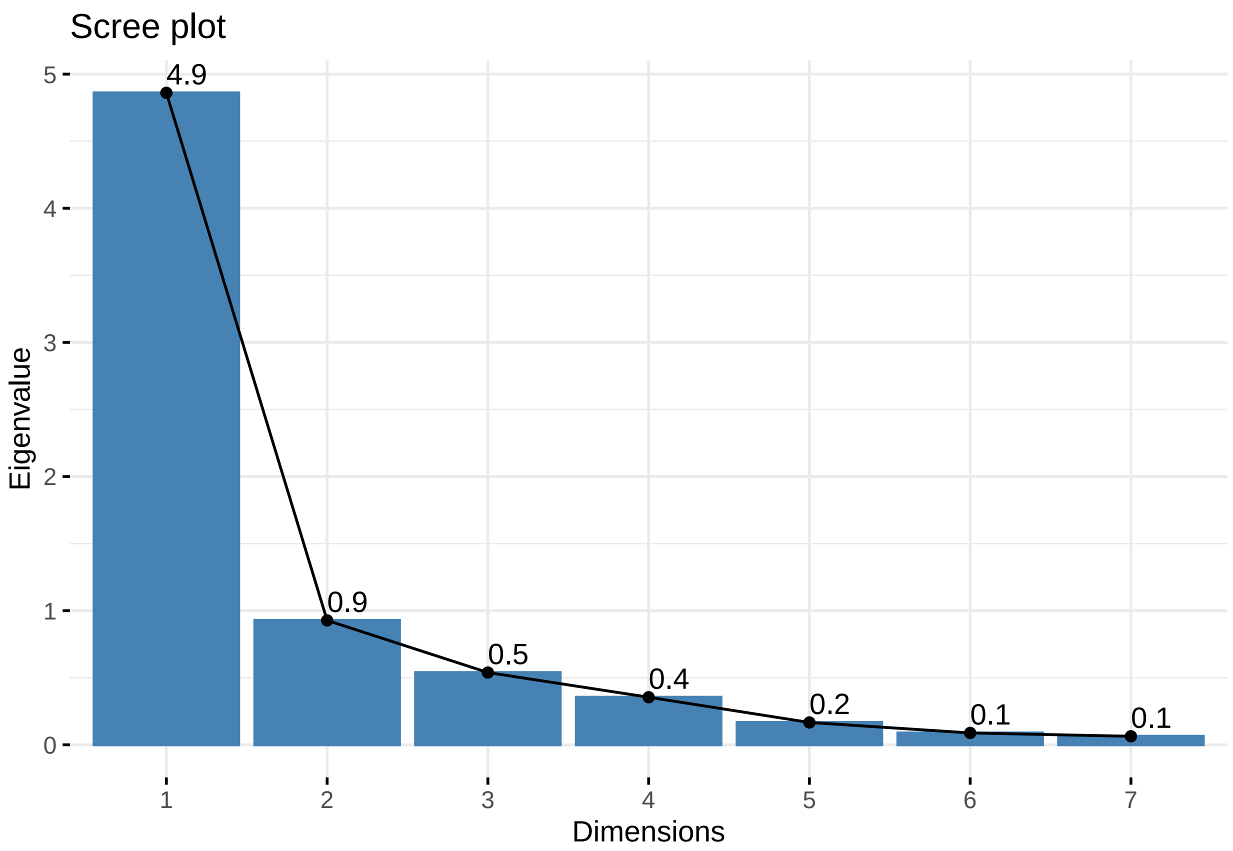
In the next PCA model I wanted to see what variables do the best job at explaining the taxonomic family of the birds. The decision tree that used the same measurements did a very poor job at classifying and the goal is to see if the PCA model provides any insights.

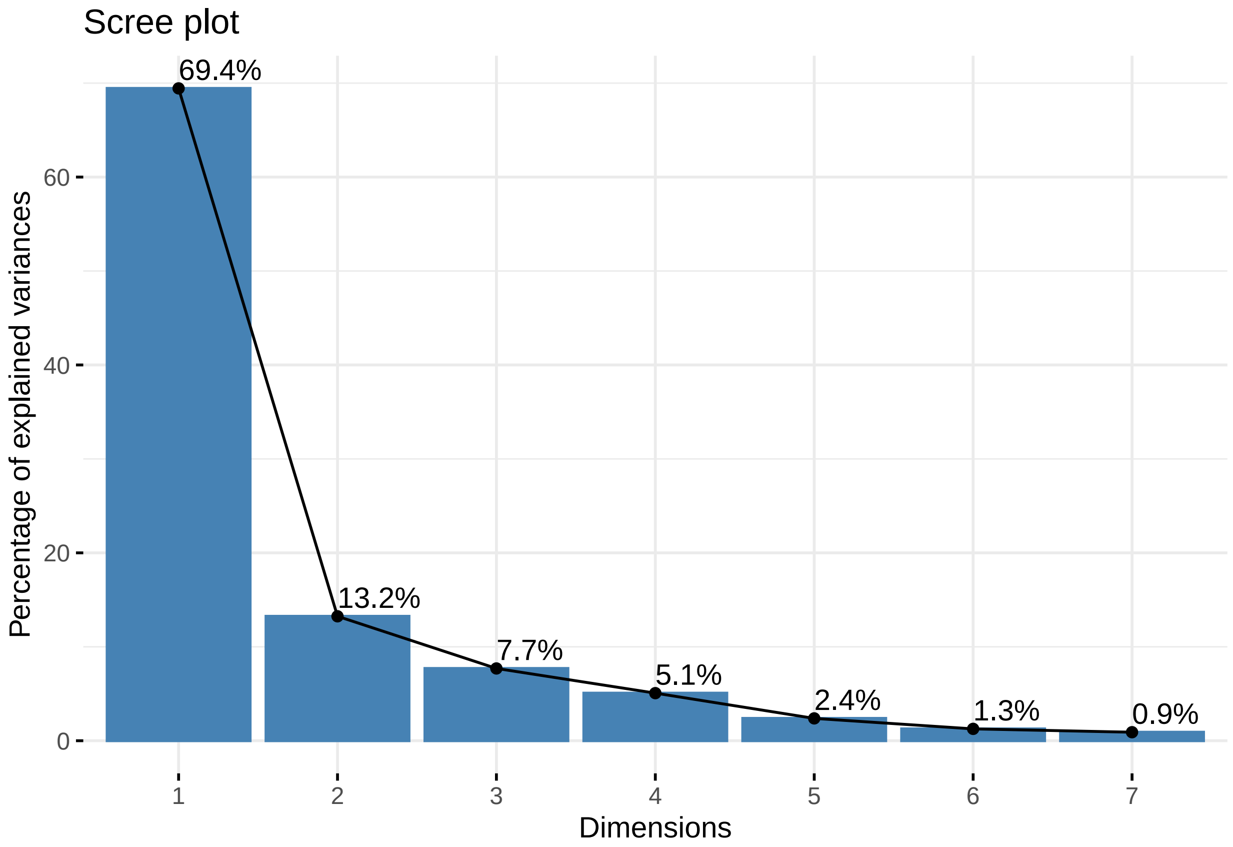
```
## [1] "Here are the eigenvectors for the 6 PCS"
```

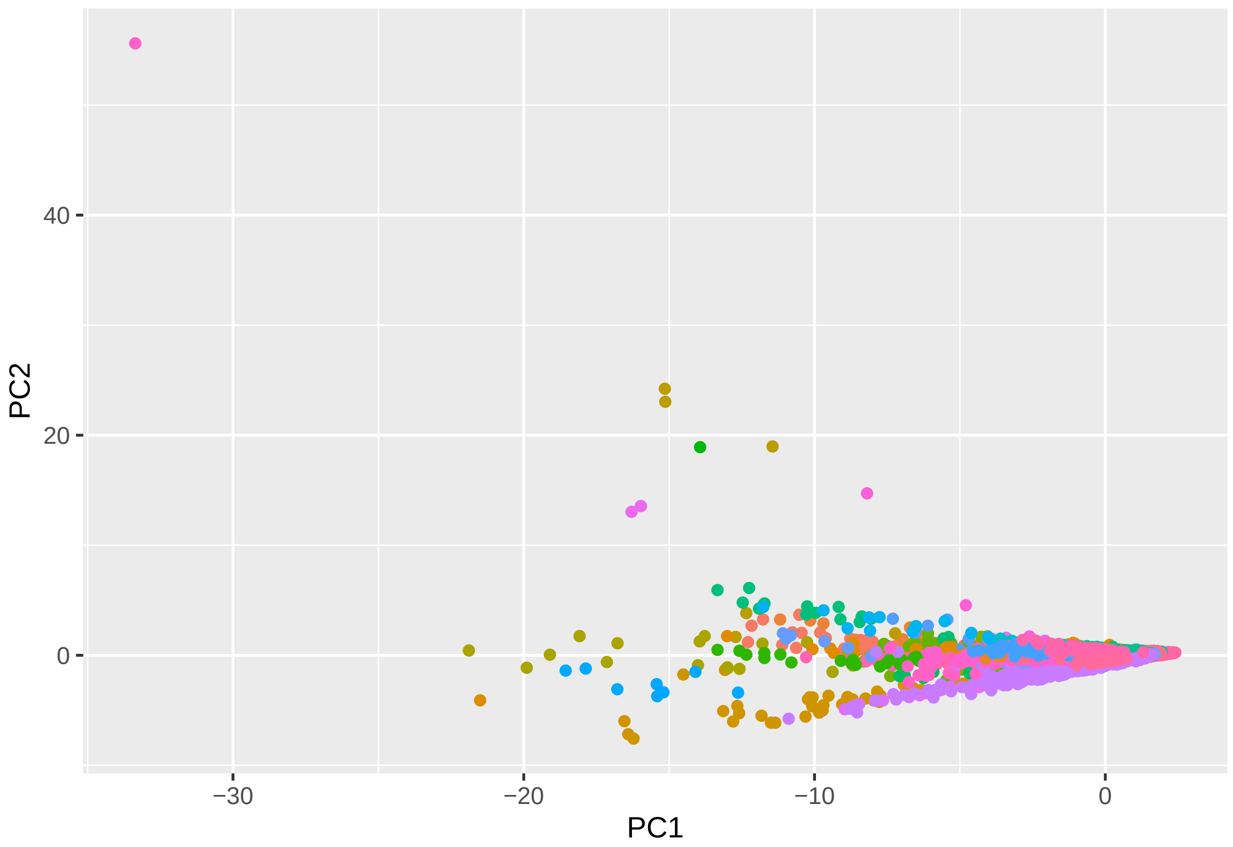
```
##                          PC1          PC2         PC3        PC4        PC5
## BeakLengthCulmen -0.3823369 -0.157589300 -0.08244397  0.8260899 -0.3355261
## BeakWidth        -0.3965389 -0.314655471 -0.39773990 -0.1496793  0.1133827
## BeakDepth        -0.3863541 -0.388163074 -0.34319008 -0.2129064  0.3178041
## TarsusLength     -0.3777435  0.373227303  0.36735809  0.2546940  0.6858350
## WingLength       -0.4130537 -0.009547704  0.38869108 -0.2617944 -0.5036244
## Secondary        -0.4184213  0.029275765  0.36895096 -0.3213924 -0.1012990
## Mass             -0.2427468  0.765009887 -0.54320255 -0.1146471 -0.1982383
##                           PC6         PC7
## BeakLengthCulmen  0.086658558  0.14026762
## BeakWidth        -0.739502186 -0.05855563
## BeakDepth         0.655201823 -0.08162394
## TarsusLength     -0.084252999 -0.20179682
## WingLength        0.039410908 -0.59538868
## Secondary          0.002980755  0.75787753
## Mass               0.087616836  0.02541205

## [1] "And here are the square roots of the eigenvalues"

## [1] 2.2047099 0.9629784 0.7339307 0.5956179 0.4080572 0.2973378 0.2521739

## [1] "These are the loadings"

## # A tibble: 7 x 6
##      ...1     ...2    ...3    ...4    ...5      ...6
##     <dbl>    <dbl>   <dbl>   <dbl>   <dbl>     <dbl>
## 1 -0.843 -0.152   -0.0605  0.492  -0.137   0.0258
## 2 -0.874 -0.303   -0.292  -0.0892  0.0463 -0.220
## 3 -0.852 -0.374   -0.252  -0.127   0.130   0.195
## 4 -0.833  0.359    0.270   0.152   0.280  -0.0251
## 5 -0.911 -0.00919  0.285  -0.156  -0.206   0.0117
## 6 -0.922  0.0282   0.271  -0.191  -0.0413  0.000886
## 7 -0.535  0.737   -0.399  -0.0683 -0.0809  0.0261
```




Variables − PCA
Mass
TarsusLength
Secondary
WingLength
akLengthCulmen
BeakWidth
BeakDepth
1.0
0.5
0.0
−0.5
−1.0
−1.0
−0.5
0.0
0.5
1.0
Dim2 (13.2%)
Dim1 (69.4%)


Scree plot
Eigenvalue
Dimensions
0
1
2
3
4
5
1
2
3
4
5
6
7
4.9
0.9
0.5
0.4
0.2
0.1
0.1


Scree plot
Percentage of explained variances
Dimensions
0
20
40
60
1
2
3
4
5
6
7
69.4%
13.2%
7.7%
5.1%
2.4%
1.3%
0.9%



Similarly with this PCA model, the overall variance explained by the first two components is very high, with this model being just a bit higher than 80%. This is interesting to see because the decision tree that tested these values produced a very bad model. In future models I would like to test to see if there is a classification algorithm that can better classify the family of the bird. With this model, like the last PCA model, the beak measurements do not do as good of a job and the other measurements provided do better.

In summary the hypothesis I made was correct, the models showed that the trophic niche and family of the birds were better explained by the PCA models than when used in a classification algorithm. Like mentioned earlier, this is not a good comparison since the models cover completely different things and PCA is for dimensionality reduction.

Summary

In this report I mainly covered the trophic niche and the family of the birds since they seemed like the more interesting one of the categorical variables. When using kNN the model did okay at classifying the niche of the bird, and did better when the habitat and wing measurements were included in the model. For the decision tree model, both models did poorly at classifying the taxonomic family of the tree. Finally, in the PCA models it was found that the measurements account for 80% or more of the variance when using the first two components. The PCA models provide some insight as to how important the measurements were and how the classification models were not suitable.

In future studies using this data set I would want to look at more unsupervised learning methods for predicting the niche and family since the supervised methods are either too limited in what they can do, or cannot handle large data. I would especially be interested to see how an artificial neural network could handle the classifications of these metrics.