Controlling wildlife trade

Muhammad Ezzat - AUG

12/14/2019

Inroduction

Undoubtedly human beings achieved culture by a lot of means, unfortunately they forgot one thing. Planet Earth, the green planet we’re living on. The number of forest fires is increasing, global warming is tearing the Earth apart & a lot of animal species are endangered. In an approach to achieve sustainability & save endangered species using modern-day technology, I’m suggesting a model that can guide us to control wildlife trade through data collected from CITES & IUCN organizations.

Importing & Joining Data

After importing data from CSV files, an outer left join is used to obtain complete observations of wildlife external trade records across the whole world. One record contains the trade details, animal traded & it’s IUCN status [ Red List ].

dim(manupilation\_df)

## [1] 92669 30

Data Manupilation

Here some data manipulation took place, the resulting data frame contains columns of Trade Appendix, Trade Term, Reported Quantity & The IUCN red list status. The goal is to build a model that classifies or predict the effect of one trade on the traded animal’s conservation status based on the first three attributes mentioned.

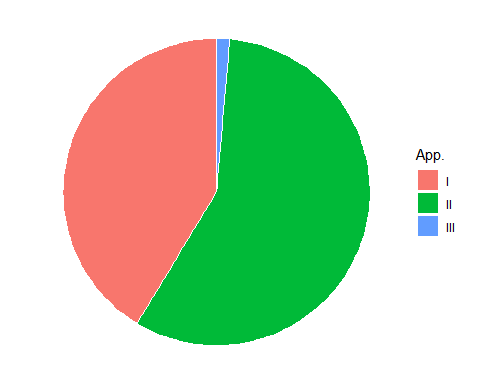
head(work\_df)

## App. Term Conservation.Status Reported.Quantity  
## 1 II live Species of Concern 2  
## 2 II live Species of Concern 2  
## 3 II live Species of Concern 2  
## 4 II live Species of Concern 2  
## 5 II live Species of Concern 2  
## 6 II live Safe 2

Exploratory Data Analysis

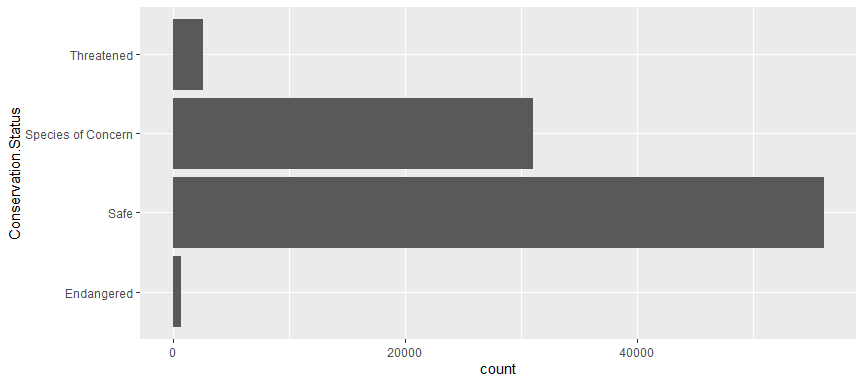
Now let's explore our data. The first graph shown in our EDA is a pie chart of the distribution of the variable Appindex , the appendix of the trade is the degree of safety of the transfer to the product delivered . It comes useful when the trade is an animal in it's living form.

ggplot(work\_df,aes(x=1,fill=App.))+geom\_bar(color="white")+coord\_polar(theta="y") + theme\_void()



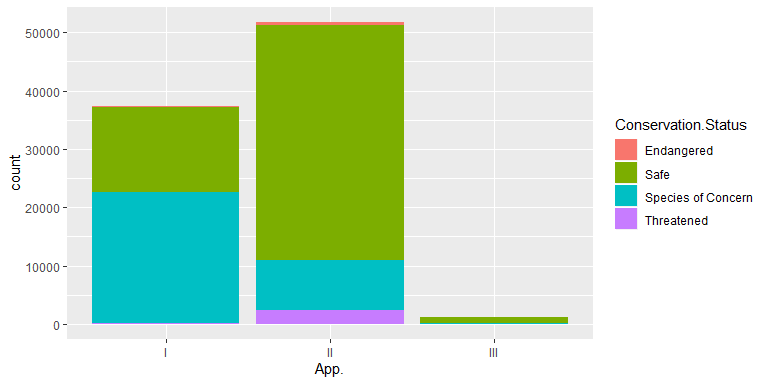
The next one is a histogram counting the number of cases in each IUCN conservation status class, where we can observe that safe species are the most ( good thing ), but the number of trades of species of concern is relatively high. The goal of the whole study is to guide that class towards the safe species class instead of being driven towards danger or threat.

ggplot(work\_df,aes(x=Conservation.Status))+geom\_histogram(stat="count")+coord\_flip()



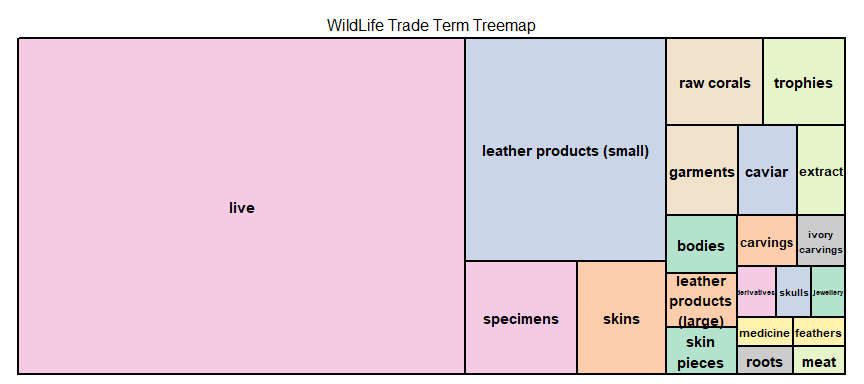
After checking on both appendix & conservation status , it's time to have a look at the combination of both variables. The proportion of conservation statues classes among each trade appendix class. It helps in understanding the distribution we have via a beautiful stacked bar chart.

ggplot(work\_df,aes(x=App. , fill = Conservation.Status))+geom\_bar()



Finally, the most beautiful of them all. A tree-map. Mapping proportion of each trade term to the area of those squares appearing. It's obvious that trading live forms of animals is the most abundant trade term in our set.

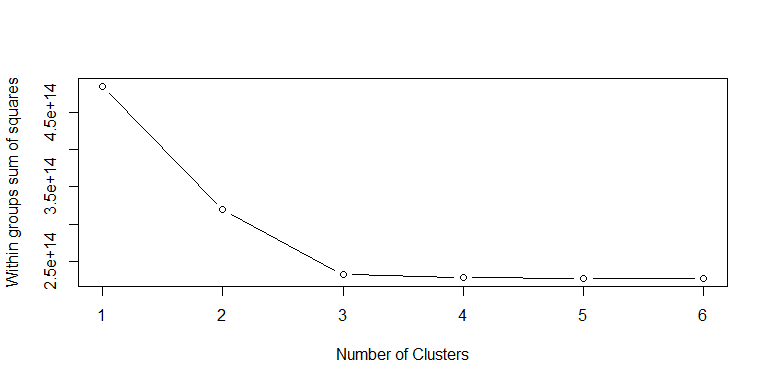
treemap(t2p, index="Term", vSize="count", type="index",   
 palette="Pastel2", title="WildLife Trade Term Treemap", fontsize.title=12)



Preparing to model

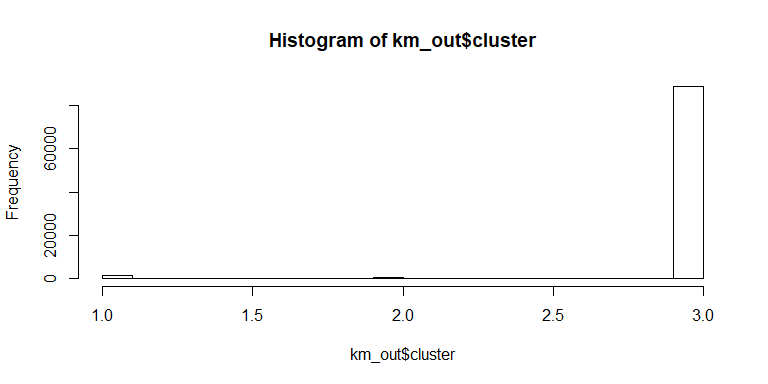
By looking at our tree-map we can observe that the variable Term had too many categories included. Regular machine learning algorithms in R can't handle them, So instead we will cluster our observations into a certain number of clusters to group them in categories according to their similarities. Statistically, that doesn't sound quite perfect, but it gives us a feasible computational solution.

plot(1:6, wss, type = "b", xlab = "Number of Clusters",   
 ylab = "Within groups sum of squares")



Here I decided to go with the most abundant clustering method, Kmeans clustering. Kmeans clustering requires the number of desired clusters to be predefined. So we tried several values of the hyperparameter K & observed the total withness space in our clusters which allow us to pick the suitable K via the elbow plot technique. Which yielded to K = 3.

km\_out <- kmeans(clustering\_df,centers=3,nstart = 20)  
hist(km\_out$cluster)



table(km\_out$cluster)

##   
## 1 2 3   
## 1265 291 88889

Modeling

Finally, it's time to build our model. Our model would simply predict the effect of the trade on the species traded based on the trade appendix, reported traded quantity & our generated cluster analysis.

control <- trainControl(method="repeatedcv", number=10, repeats=3)  
set.seed(7)  
fit.rf <- train(Conservation.Status ~ App. + Reported.Quantity + cluster,  
 data= train\_data, method="rf", trControl=control)

rf\_predictions <- predict(fit.rf,test\_data)  
test\_data$Conservation.Status <- as.factor(test\_data$Conservation.Status)  
m <- confusionMatrix(rf\_predictions,test\_data$Conservation.Status)  
print(m$overall)

## Accuracy Kappa AccuracyLower AccuracyUpper AccuracyNull   
## 7.104064e-01 4.302109e-01 7.044455e-01 7.163132e-01 6.214232e-01   
## AccuracyPValue McnemarPValue   
## 4.298488e-174 4.701843e-281

Conclusion

On assessing the model’s quality we yielded an accuracy of 70 % which is quite satisfying to me. As George Box once said “All models are wrong, but some are useful”, I find the idea of this model so attractive that we can apply later on with richer data that contain more descriptive attributes of an animal, that can be a good step towards achieving environmental sustainability.