Report on CMM535

2/19/2022

#Report on CMM 535 ## Introduction I had to make use of an external tool apart from R due to my familiarity with the platforms used. This made it easy to extract the needed data from the convoluted dataset. I used a website(Regex101.com) to implement regular expression for string manipulation and data extraction. This was done for some columns(Age, average weight and final reason for escape) contained within the escape.csv. This decision was purely made due to the fact that I was more familiar with using that platform for my data cleaning process. I had to set my current working directory using the sessions tab after the data cleaning process in order to have complete access to the data I was working with.

#Ensure the current working directory is set to the path where your script and csv files are located  
setwd("C:/Users/DELL/Desktop/Data ass")

This also created a priority and preference when it came to referring to the data sets used for this analysis. Prior to starting, I had to load the tidyverse package so that I could manipulate the clean data in a data frame and further clean it in preparation for future analysis in order to extract some key insights from our data sets.

**import the required libraries** It is necessary to require or import the tidyverse package into your workflow after the initial preparation in order for the project to continue smoothly.

library(tidyverse)

## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --

## v ggplot2 3.3.5 v purrr 0.3.4  
## v tibble 3.1.6 v dplyr 1.0.8  
## v tidyr 1.2.0 v stringr 1.4.0  
## v readr 2.1.2 v forcats 0.5.1

## -- Conflicts ------------------------------------------ tidyverse\_conflicts() --  
## x dplyr::filter() masks stats::filter()  
## x dplyr::lag() masks stats::lag()

The tidyverse package is a very rebust data wrangling tool that can be deployed on R. It’s worth noting that the first installation of the library required for the analysis can cause some issues. This is because filter, for example, is already installed on the base requirement of R, and that may be obscured by the new library’s installation.

From this point forward, we’ve completed our first step of installing the necessary libraries so that we may move on to the core analysis. **From the escape.csv, we shall import the first batch of data** We then proceed to import our data sets into their relevant variables after loading the required packages that will be utilized for data wrangling and additional data cleaning of our data set.

escape <- read.csv("escapes CMM 535 Coursework - clean.csv",header=TRUE,sep=",",stringsAsFactors = T)

## Data cleaning with R

The first order of business was to explore the individual columns that were meant to be treated and dealing with outlier values, data type consistency and missing values. We will start with the Age column and then proceed from there.

# get a sense of the range and distribution of the values contained here.  
summary(escape$Age)

## Min. 1st Qu. Median Mean 3rd Qu. Max. NA's   
## 1.0 11.0 14.0 242.5 21.0 2019.0 43

From the above result, we can clearly see that we have some outliers within the Age column. This was cited when we noticed that the max value of the column is 2019. With simple common sense, we can tell that is impossible. Our job would therefore be to normalize it by removing the outliers and using the mean of the ages to replace NA values The code below achieved that.

#We have to take into consideration that there are some outliers in the dataset imported   
  
escape$Age[escape$Age > 100] <- NA #remove outlier  
escape$Age[is.na(escape$Age)] <- mean(escape$Age,na.rm = T)

While exploring my data collection, I also went one step further and changed the format of my data types from factor to their appropriate integer kinds.

# We then proceed to the next column of interest, which is the   
#Deal with   
escape$Average.Weight.Kg. <- as.numeric(as.character(escape$Average.Weight.Kg.))

## Warning: NAs introduced by coercion

escape <- escape[!is.na(escape$Average.Weight.Kg.),]  
escape$Average.Weight.Kg.[escape$Average.Weight.Kg. > 50] <- NA  
escape$Average.Weight.Kg.[is.na(escape$Average.Weight.Kg.)] <- mean(escape$Average.Weight.Kg.,na.rm = T)  
  
str(escape$Final.Number.Escaped)

## Factor w/ 182 levels "#N/A","0","1",..: 123 121 3 1 99 109 2 3 69 149 ...

str(escape$Final.Number.Recovered)

## Factor w/ 37 levels "#N/A","0","1",..: 2 2 2 1 2 2 2 2 2 2 ...

#Fix the last numerical columns that will be used for typing  
escape$Final.Number.Escaped <- as.integer(as.character(escape$Final.Number.Escaped))

## Warning: NAs introduced by coercion

escape$Final.Number.Recovered <- as.integer(as.character(escape$Final.Number.Recovered))

## Warning: NAs introduced by coercion

escape$Final.Number.Escaped[is.na(escape$Final.Number.Escaped)] <- 0  
escape$Final.Number.Recovered[is.na(escape$Final.Number.Recovered)] <- 0

The values that had no distinguishable values were converted to NA values. Another thing that must be considered while considering the columns that must be used for the analysis, is editing the unique names of the final reasons for fishes escaping. It would be a desirable approach to change the default values in the column of interest.

#First let us get a sense of the unique values that are contained in the column.  
unique(escape$Final.Escape.Reason)

## [1] other hole in net   
## [3] human error <NA>   
## [5] weather no actual escape of fish   
## [7] predator equipment damage   
## [9] screen failure equipment wear and tear   
## [11] net failure (not including hole) vandalism   
## [13] transfer pipe failure chafe/ snag   
## [15] inappropriate equipment mooring failure   
## [17] flooding unknown   
## [19] pen failure   
## 18 Levels: chafe/ snag equipment damage equipment wear and tear ... weather

#Something that was unique about this is the fact that there was a reason that stated "no actual escape of fish"  
#Going forward, we should exclude that from our query through a filter function  
realEsc <- escape %>%   
 replace\_na(list(Final.Escape.Reason = "unknown")) %>%   
 filter(Final.Escape.Reason != "no actual escape of fish") %>%  
 mutate(Final.Escape.Reason = recode(Final.Escape.Reason,  
 `other` = "unknown"))

Notice that I had to group a lot of the initial values as “unknown” since they all fall within the same category. Moving on, we’ve prepared our data for analysis, we can move on to the following step. We will now attempt to clean up some entries as they related

# Exploratory Data Analysis using some descriptive statistical analysis

## Exploratory Data Anlysis

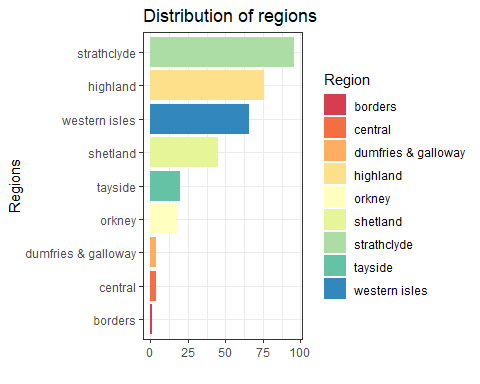
We’ll need to import the ‘ggplot’ package because we’ll be doing some data visualization. ‘Ggplot’ is a powerful data visualization software.

#Exploratory Data Analysis using some descriptive statistical analysis  
library(ggplot2)

**Background for analysis** We want to embark on the journey of asking some pertinent questions about the dataset given to us. Asking these questions about the data sets also requires us to think deeply about how they influence the overall result. There will also be visualizations attached to each of the questions in order to easily understand the insights derived.

**Q1 What were the major regions had the most frequency for our dataset??** We will attempt to find the distribution of the escaped species below, and visualize the using the “ggplot” library.

region <- escape %>%   
 count(Region) %>%   
 arrange(desc(n))  
   
region %>%   
 ggplot(aes(x=reorder(Region,n),y=n, fill = Region))+  
 geom\_col() + coord\_flip() +  
 scale\_fill\_brewer(palette="Spectral")+  
 theme\_bw()+  
 ggtitle("Distribution of regions")+  
 xlab("Regions") + ylab("")



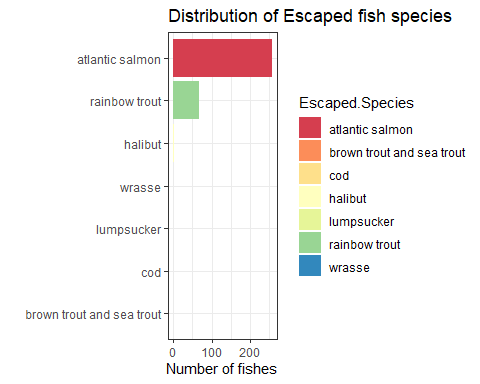
#Going forward, the top 5 should be of heavy interest

The strathclyde, highland, and western isles regions are the top three spots that allowed fish to escape, according to the above visualization. Further investigation is needed to gain a better knowledge of their procedures that allow fish to be lost. But in essence, we have to pay further attention to those locations going forward.

**Q2 What fish specie had the most number of escape?**

Another question to ask is about the prevalent reasons for the loss of the fishes. The code below can answer that question.

fish <- escape %>%   
 count(Escaped.Species) %>%   
 arrange(desc(n))  
  
fish %>%   
 ggplot(aes(x=reorder(Escaped.Species,n),y=n, fill = Escaped.Species))+  
 geom\_col() + coord\_flip() +  
 scale\_fill\_brewer(palette="Spectral")+  
 theme\_bw()+  
 ggtitle("Distribution of Escaped fish species")+  
 xlab("") + ylab("Number of fishes")

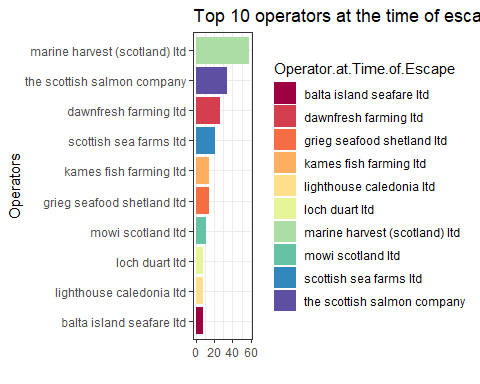


# create a new data frame that filters the essential values to be analyzed  
primaryFish <- escape %>%   
 filter(Escaped.Species == "atlantic salmon" |  
 Escaped.Species == "rainbow trout")

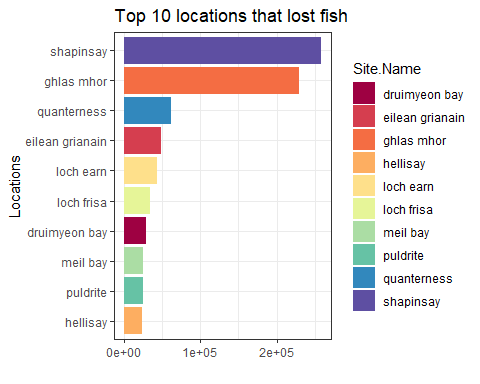
As illustrated in the graph above, the Atlantic salmon and rainbow trout are two species worth studying. These two categories have a wide variation over the Escaped species variable in our sample. As a result, before we continue our research, we should filter those categories so that we may conduct further analysis due to the minimal influence of the other fish species.

**Q3 Discussion about the operators and location at the time of loss** The reason for this question is so that we can get a sense of perspective about the operators and location with a high risk of losing the fishes. We’ll also keep running a distribution query to figure out which sites in our data sets have the best distribution. This will give us an idea of what the location with the most loss is like.

locat <- escape %>%  
 drop\_na(Operator.at.Time.of.Escape) %>%   
 select(Operator.at.Time.of.Escape) %>%  
 count(Operator.at.Time.of.Escape) %>%   
 arrange(desc(n)) %>%   
 slice(1:10)  
  
locat %>%   
 ggplot(aes(x=reorder(Operator.at.Time.of.Escape,n),y=n, fill = Operator.at.Time.of.Escape))+  
 geom\_col() + coord\_flip() +  
 scale\_fill\_brewer(palette="Spectral")+  
 theme\_bw()+  
 ggtitle("Top 10 operators at the time of escape")+  
 xlab("Operators") + ylab("")



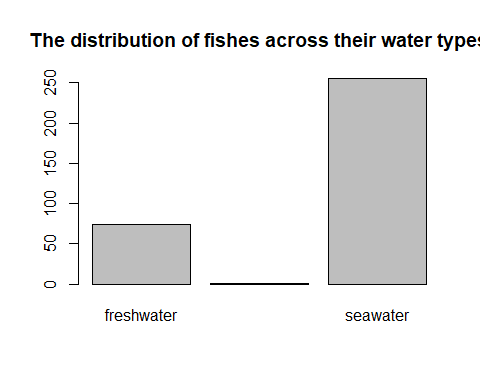
#top 10 locations that were responsible for losing fish  
locatLoss <- escape %>%   
 filter(Final.Number.Escaped > 0) %>%   
 group\_by(Site.Name) %>%   
 summarise(total.Loss = sum(Final.Number.Escaped)) %>%   
 arrange(-total.Loss) %>%   
 slice(1:10)  
  
  
   
locatLoss %>%   
 ggplot(aes(x=reorder(Site.Name,total.Loss),y=total.Loss, fill = Site.Name))+  
 geom\_col() + coord\_flip() +  
 scale\_fill\_brewer(palette="Spectral")+  
 theme\_bw()+  
 ggtitle("Top 10 locations that lost fish")+  
 xlab("Locations") + ylab("")

 We can plainly observe the operators who were responsible for the most to the least significant numbers of losses from the chart above. **marine harvest (Scotland), Scottish Salmon company, and dawnfresh** are some of the notable operators to keep an eye on. Combining this data with other findings will provide us with a wide range of options for who would be the ideal operator at each site. In general, though, these businesses should be appreciated for their capacity to recover losses. The chart above gives a visual representation of the top 10 locations that had fish. Next up, we query the number of fishes that were lost by each Operators. This will be discussed later.

**Q4 An investigation of water type of the escaped species** *First we need to establish some fundamental ideas with the information before we proceed to our analysis and visualization*

We simply just want to know the dominant water type that the fishes were exposed to here. This gives some perspective when it comes to taking priority with this feature in mind.

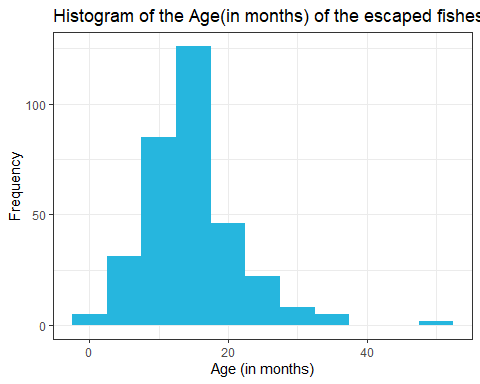
barplot(table(escape$Water.Type),main = "The distribution of fishes across their water types")



Seawater is the most dominant habitat. An interesting finding is that the majority of instances occurred at areas with seawater, with the remainder occurring at sites with a mix of both. As we move forward, we must concentrate more on the seawater cases.

**Q5 What is the age distribution and average age and average weight distribution?** In order to answer this question, there is a need to use a histogram to inspect this in order to give a clear answer on most frequent ages in months. The plot to show this can be shown below.

escape %>%   
 ggplot(aes(Age))+  
 geom\_histogram(binwidth = 5 ,fill= "#26b6de")+ #a bin size of 5 was used  
 theme\_bw()+  
 labs(title = "Histogram of the Age(in months) of the escaped fishes",  
 x = "Age (in months)",  
 y = "Frequency")



What we can see from the histogram distribution above is that the great bulk of the age distribution falls between 15 and 25 months (since the bin was 5).

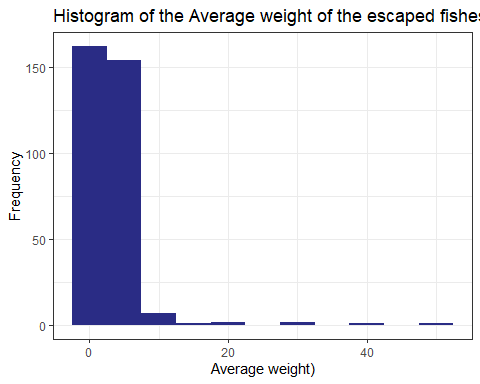
For the average of the average weight distribution, a simple query can resolve it.

mean(escape$Average.Weight.Kg.)

## [1] 3.05018

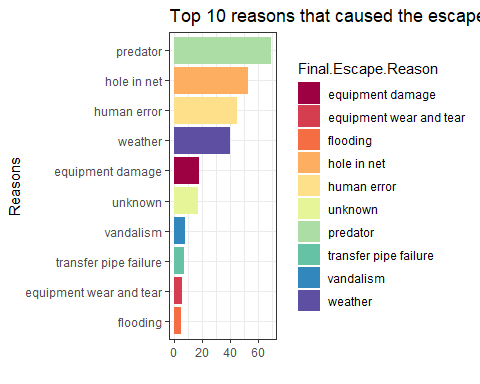
The average weight is apparently 3.05018 kg.

escape %>%   
 ggplot(aes(Average.Weight.Kg.))+  
 geom\_histogram(binwidth = 5 ,fill= "#2a2c85")+   
 theme\_bw()+  
 labs(title = "Histogram of the Average weight of the escaped fishes",  
 x = "Average weight)",  
 y = "Frequency")

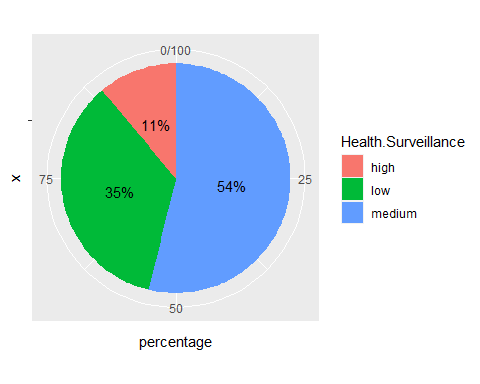


**Q6 What were the most common reasons for the loss of fishes?** Now to investigate what the distribution of escaped fishes are across the various water types. We will show this on a barplot of ease of viewing this insight.

freason <- realEsc %>%  
 drop\_na(Final.Escape.Reason) %>%   
 select(Final.Escape.Reason) %>%  
 count(Final.Escape.Reason) %>%   
 arrange(desc(n)) %>%   
 slice(1:10)   
  
#Now we continue with the visualization of the data  
  
  
  
freason %>%   
 ggplot(aes(x=reorder(Final.Escape.Reason,n),y=n, fill = Final.Escape.Reason))+  
 geom\_col() + coord\_flip() +  
 scale\_fill\_brewer(palette="Spectral")+  
 theme\_bw()+  
 ggtitle("Top 10 reasons that caused the escape of the fishes")+  
 xlab("Reasons") + ylab("")

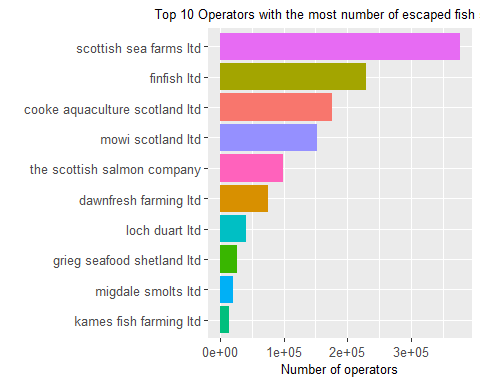
 Predators, holes and human error are the three most prevalent reasons for fishes being lost from their location, as displayed in the graph above. There also appears to be a significant proportion of the loss attributed to genuine fish escape. With the aforementioned factors in mind, this suggests that more inquiry is required in order to reduce the danger of future losses.

**Q7 What the representation of health for the fishes considered?** We also wanted to understand what the health surveillance level of the various sites and fishes were. This gives us a proper perspective of what to expect. A pie chart would be ideal for visualizing the answer to this question.

 Because the data shows that fish with high health surveillance have the lowest representation, it is clear that a high safety standard is critical to the fishes’ survival. We can acquire an idea of the top operators who are commendable and the conditions that are primarily responsible for the loss of our fishes based on what we can tell about the data set.

**Q8 What operator had the largest number of lost fishes?**

# A look at operators that loss the most amount of fishes  
opCum <- escape %>%   
 group\_by(Operator) %>%   
 summarise(cum = sum(Final.Number.Escaped)) %>%   
 arrange(desc(cum)) %>%   
 slice(1:10)  
  
ggplot(data = opCum) +  
 geom\_bar(aes(x = reorder(Operator, cum), y = cum, fill = Operator),  
 stat = "identity",show.legend = F) +  
 #theme\_bw() +  
 ggtitle("Top 10 Operators with the most number of escaped fish species") +  
 theme(plot.title = element\_text(hjust = 0.5)) +  
 theme(axis.text = element\_text(size = 10),  
 axis.title = element\_text(size = 10),  
 plot.title = element\_text(size = 10)) +  
 xlab("") + ylab("Number of operators") +  
 coord\_flip()

 Scottish sea farms, finfish ltd and cooke aquaculture scotland ltd are the major operators that are responsible for the escape of fishes on the farms. They should be released from their duties.

#Merging the two given datasets with the “Site name” variable The last portion of the project required us to combine the data from escape.csv and analysis.csv into a single csv file called “escapesPlus.csv.” We have to use the “merge” function in order to accomplish this. Because left, right, and inner joins didn’t account for a lot of values, a full join was performed. The “Site.Name” column was chosen as the common column for this merging. Before we begin the merge, we have to import the second data set that will be used.

analysis <- read.csv("analysis CMM 535 coursework.csv",header = T,stringsAsFactors = T)  
  
escapePlus <- merge(escape,analysis,by = "Site.Name", all.x = T, all.y = T)  
write.csv(escapePlus, file="escapesPlus.csv", row.names = F)

Another justification for the full based on the requirements of this assignment, is due to the fact there was no corresponding primary column to be utilized. And with taht we’ve concluded our analysis for this assignment. The insights shared through visualizations can be quite helpful to tackling the problem of increased escaping of fishes.