Table of content

1713103_Part1

AFatima

14/03/2022

Overview

We are using two data sets from Aquaculture (fish farms) to discuss main concepts and tools for a data science project, escapes.csv that contains records of fish escapes and incidents and analysis.csv that contains the results of water analysis using several components. This has aligned records with escape.csv. Which will be later used for merging both. We are loading, exploring, modelling and visualizing data using off-the-shelf tools and packages in R. We have two datasets from Aquaculture (fish farms), where fish are grown in large cages, either in the sea or in lakes. Record keeping is required for monitoring the fish farms for any incidents of escapes for the specific fishes that has been kept and monitored in cages. Also, in the escapes.csv our focus is Escaped Species, Age and the Average weight.

Loading libraries

```
library(dplyr) ## For data preparation
library(caret) ## For value imputation
library(Hmisc)
library(lubridate) ## For time intelligence
library(stringr)
library(tidyverse)
library(data.table)
library(corrplot)
library(lattice)
library(rattle)
library(randomForest)
library(ggplot2)
```

Loading datasets

```
escapes <- read.csv("escapes.csv", header = T, stringsAsFactors = T)
analysis <- read.csv("analysis.csv", header = T, stringsAsFactors = T)
```

TASK 1

Exploring and cleaning features of the escapes dataset:

- Checking summary of data set to see descriptive statistics and distribution for focussed features of data set escapes. Escaped species, avg weight and age with uni variate analysis/statistics, include missing data and outliers, visualizing individual focused features using plots and histograms. Checking whether distribution is normal or skewed.
- Bi variate analysis of target feature relationship with other features. Identifying redundant variables and deleting them. using visualization boxplot for nominal vs numeric variables, scatter plots for pairs of numeric variables, tables for pairs of nominal(categorical) variables
- In the task we noticed we have Unknown values in escape start time, Average weight, Initial number escaped, final number escaped attributes
- There are NA's in Escape.Start.Time, Escape.End.Time, Escape.Grid.Reference, Age, Average.Weight, Initial.Date.of.Escape, Initial.Number.Escaped,Final.Date.of.Escape, Final.Number.Escaped, Final.Number.Recovered, Final.Escape.Reason, Site.Address.1, Site.Address.2, Site.Post.Code, Site.Contact.Number, Health.Surveillance and MS.Management.Area.
- The data set has dimensions 357 rows and 38 columns.Data cleaning has been done for each feature at a time while exploring. Plotting bar plot frequency counts for escaped species we can clearly see the "Atlantic salmon" and "rainbow trout" has much more escaped incidents when compared with any other species. -"Age" column required data cleaning with string processing functions such as str_remove_all and function to replace ranges by mean, data type conversion, renaming column so that it represents the age in months. Histogram and quantiles has been plotted which shows the attribute is distributed about 25% in 12 months old species, while 50% 15 months and maximum is 48 month old.
- Same process has been applied for Avg weight attribute cleaning with str_remove_all function and then conversion of kgs to grams by passing if else and locating string via grep1. The plot shows outliers. Hence, Inter quartile range 600, range 1-10000, variance 1496411 standard deviation 1223.279 has been obtained showing the distribution.
- From two attributes of dates in the data set Escape.start.date is the one that has matching and aligning records with the data set analysis.csv. So, I am choosing this specific date column and discarding the other in next steps. Attribute is parsed to take via "ymd" function from "lubridate" package. Splitting Date column into three separate columns as Date, Month and Year. Removing "0" from months column to match with analysis.csv and merging year and month for task 2 by using paste function.
- Deleting unimportant attributes. Operator at time of escape has been prepossessed to delete "ltd." in strings. Final number escaped string reprocessing has been done with str_remove_all function and noise allocated to "NA" then imputing mean values for ranges using sapply and converted to numeric.

- Final escape reason converted to datatype factor from character. There was 5 NA's which later has been imputed to random values using mode.
- Marine.Scotland.site.ID doesn't have any NA's or duplicates so no cleaning required it has 190 levels
- Site name has 189 levels and no duplicates or cleaning required
- Producing.in.Last.3.Years has two levels "yes" is highly distributed than "no" in the attribute
- water type has 3 levels

```
summary(escapes)

##    Escape.ID                      Operator.at.Time.of.Escape
Escape.Water.Type
##  Min.   :2000001   marine harvest (scotland) ltd: 59          b:  1
##  1st Qu.:2000153   the scottish salmon company  : 36          f: 85
##  Median :2000295   dawnfresh farming ltd        : 27          s:271
##  Mean   :2000289   scottish sea farms ltd       : 22
##  3rd Qu.:2000417   kames fish farming ltd       : 18
##  Max.   :2000527   grieg seafood shetland ltd   : 17
##                    (Other)                      :178
##  Escape.Start.Date Escape.Start.Time  Escape.End.Time
Escape.Grid.Reference
##  11-Jan-05:  8     12:00  : 24       12-Jan-05:  3    nn762442:  7
##  13-Sep-06:  3     10:00  : 14       18-May-09:  3    nb137355:  6
##  18-May-09:  3     unknown: 14       17-Jan-12:  2    nf759032:  5
##  29-Jan-00:  3     11:00  : 12       17-Jan-20:  2    nn620240:  5
##  01-Apr-00:  2     09:00  : 10       23-Jan-14:  2    nr816412:  5
##  01-Apr-01:  2     (Other):113       (Other)  :200    (Other) :225
##  (Other)  :336     NA's   :170       NA's     :145    NA's    :104
##                   Escaped.Species                              Stage
##  atlantic salmon          :277    broodfish                       :  3
##  brown trout and sea trout:  1    fish weighing more than 5 grams: 84
##  cod                      :  1    grower fish (salmon only)       :247
##  halibut                  :  2    salmon fresh water stages       : 23
##  lumpsucker               :  1
##  rainbow trout            : 74
##  wrasse                   :  1
##         Age      Average.Weight Initial.Date.of.Escape
Initial.Number.Escaped
##  12 months: 28   unknown: 15    14-Jan-05:  4          0        : 67
##  unknown  : 22   1 kg   : 13    13-Jan-05:  3          unknown  : 64
##  9 months : 16   3 kg   : 13    17-Jan-05:  3          1        : 12
##  15 months: 14   2.5 kg :  9    25-May-09:  3          not known:  8
##  18 months: 14   4 kg   :  9    02-Sep-19:  2          none     :  7
##  (Other)  :247   (Other):294    (Other)  :290          (Other)  :198
##  NA's     : 16   NA's   :  4    NA's     : 52          NA's     :  1
##                    Initial.Escape.Reason Final.Date.of.Escape
##  hole in net - hol             :68        04-Sep-19:  4
##  predator - prd                :65        03-Sep-18:  3
##  human error - hum             :49        08-Feb-05:  3
```

```
##  no actual escape of fish - nes:48        09-Feb-05:  3
##  weather - wth                  :47        25-Mar-20:  3
##  equipment damage - eqd         :21        (Other)  :283
##  (Other)                        :59        NA's     : 58
##  Final.Number.Escaped Final.Number.Recovered
##  0      : 98          0       :205
##  1      : 12          n/a     : 31
##  200    :  6          none    :  6
##  unknown:  6          1       :  4
##  20000  :  5          80 - 100:  3
##  (Other):225          (Other) : 46
##  NA's   :  5          NA's    : 62
##                      Final.Escape.Reason Marine.Scotland.Site.ID
##  predator - prd                 :72       fs0180 :  9
##  hole in net - hol              :57       fs0432 :  8
##  no actual escape of fish - nes:54        fs0717 :  8
##  human error - hum              :49       fs1176 :  8
##  weather - wth                  :44       fs0150 :  6
##  (Other)                        :76       fs0268 :  6
##  NA's                           : 5       (Other):312
##   Date.Registered           Site.Name   National.Grid.Reference
##  01-Sep-84:  9    loch earn      :  9   nn620240:  9
##  16-Oct-91:  9    balta isle     :  8   hp657082:  8
##  23-Mar-10:  9    eilean grianain:  8   nn762442:  8
##  01-Jul-80:  8    loch tay       :  8   nr816412:  8
##  01-Jan-88:  7    loch lochy     :  6   nb126343:  6
##  01-Jan-78:  6    taranaish      :  6   nb177372:  6
##  (Other)  :309    (Other)        :312   (Other) :312
##           Local.Authority Producing.in.Last.3.Years
##  argyll and bute  :98     no : 53
##  highland         :87     yes:304
##  western isles    :71
##  shetland         :48
##  perth and kinross:24
##  orkney           :19
##  (Other)          :10
##                Site.Address.1        Site.Address.2       Site.Address.3
##  n/a                   : 38    uig           : 15    argyll       : 45
##  gremista              : 14    lerwick       : 14    shetland     : 45
##  loch etive trout farm: 13    inverawe      : 13    isle of lewis: 23
##  miavaig pier          : 12    unst          : 12    n/a          : 21
##  uyeasound             : 12    scourie, lairg: 11    ross-shire   : 15
##  badcall salmon house : 11    (Other)       :254    (Other)      :170
##  (Other)               :257    NA's          : 38    NA's         : 38
##   Site.Post.Code   Site.Contact.Number Aquaculture.Type
##  ze1 0px : 14    01546 602172: 67      fish:357
##  pa35 1hu: 13    01856 876101: 18
##  hs2 9hw : 12    01397 715032: 14
##  ze2 9dl : 12    01971 502451: 14
##  iv27 4th: 11    01595 741817: 13
```

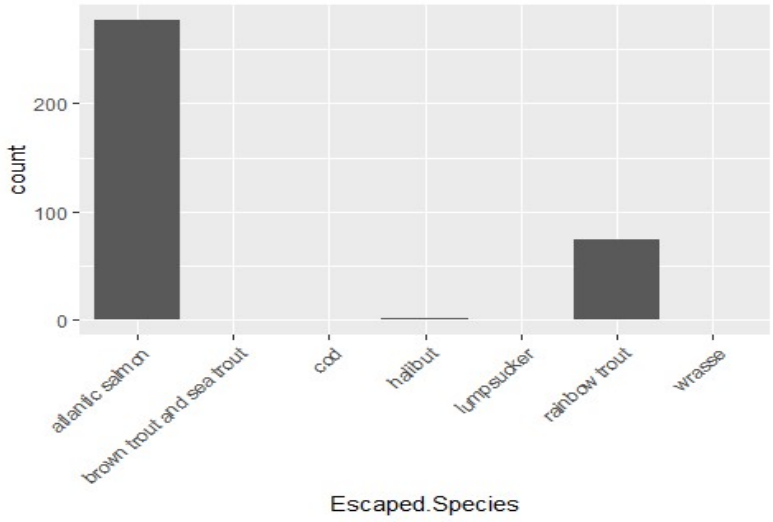
```
##  (Other) :257    (Other)     :193
##  NA's    : 38    NA's        : 38
##                    Water.Type      Health.Surveillance    Easting
##  freshwater             : 85   high          : 36      Min.   : 75200
##  freshwater and seawater:  1   low           :118      1st Qu.:143500
##  seawater               :271   medium        :187      Median :192400
##                                not applicable:  4      Mean   :222677
##                                NA's          : 12      3rd Qu.:262000
##                                                        Max.   :465700
##
##     Northing
##  Min.   : 580000
##  1st Qu.: 733100
##  Median : 836500
##  Mean   : 857400
##  3rd Qu.: 935500
##  Max.   :1208200
##
##                                                    MS.Management.Area
##  not in a management area                                      : 83
##  15b - linnhe, firth of lorne, sound of mull and loch sunart: 38
##  6a - loch roag                                                : 23
##  8b - central orkney                                           : 15
##  3a - sw shetland mainland                                     : 14
##  (Other)                                                       :155
##  NA's                                                          : 29
##            Region                                Operator
##  strathclyde  :99   mowi scotland ltd             :78
##  highland     :87   the scottish salmon company   :76
##  western isles:71   dawnfresh farming ltd         :40
##  shetland     :48   scottish sea farms ltd        :38
##  tayside      :23   cooke aquaculture scotland ltd:36
##  orkney       :19   grieg seafood shetland ltd    :25
##  (Other)      :10   (Other)                       :64
##                                  Species
##  atlantic salmon                     : 63
##  rainbow trout                       : 61
##  atlantic salmon, wrasse, lumpsucker: 40
##  atlantic salmon, lumpsucker, wrasse: 34
##  lumpsucker, atlantic salmon, wrasse: 30
##  wrasse, atlantic salmon, lumpsucker: 27
##  (Other)                             :102
```

```
# Checking dimensions
dim(escapes)
```

```
## [1] 357  38
```

Escaped.Species:

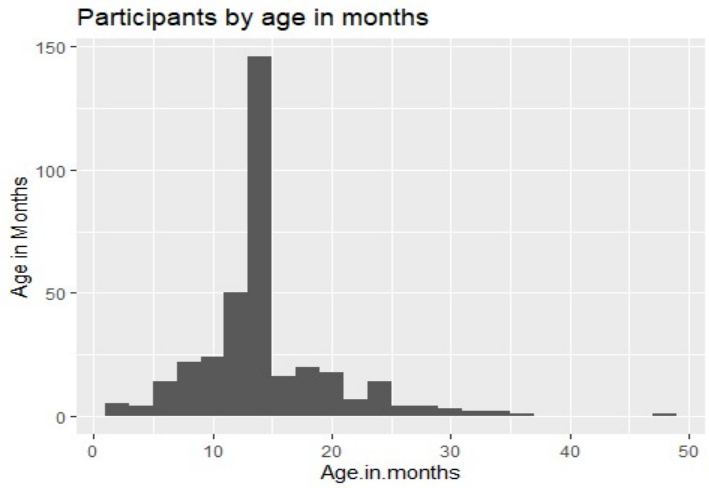
```
## bar plot
p <- ggplot(data=escapes, aes(x=Escaped.Species))
 p <- p + geom_bar()+  theme(axis.text.x = element_text(angle = 45,
                                    hjust = 1))
 p
```



Age:

```
escapes$Age <- str_remove_all(escapes$Age, " ")
escapes$Age <- str_remove_all(escapes$Age, "~")
escapes$Age <- str_remove_all(escapes$Age, " months")
escapes$Age <- str_remove_all(escapes$Age, "months")
#if there is a - sign, a range is provided
findMean <- function (svalue){
   if(grepl("-", svalue)){
# replace ranges by mean
      minus <- str_locate(svalue,"-")[1,1]
      num1 = as.numeric(substr(svalue, 0,minus-1))
      num2 = as.numeric(substr(svalue,minus+1, str_length(svalue)))
      return(mean(c(num1,num2)))
   } else {return(as.character(svalue))}
}
escapes$Age <- sapply((escapes$Age),findMean)

## Warning in FUN(X[[i]], ...): NAs introduced by coercion

## Warning in FUN(X[[i]], ...): NAs introduced by coercion

## Warning in FUN(X[[i]], ...): NAs introduced by coercion

## Warning in FUN(X[[i]], ...): NAs introduced by coercion

## Warning in FUN(X[[i]], ...): NAs introduced by coercion
```

```
escapes$Age[escapes$Age == "1999"] <- "NA"
## Converting from character to numeric
escapes$Age <- as.numeric(escapes$Age)

## Warning: NAs introduced by coercion

## Imputing mean for NA's
escapes$Age <- round(impute((escapes$Age), mean))
##renaming column name
escapes$Age <- rename(escapes, Age.in.months = Age)
## Dropping Age level
escapes <- drop(escapes$Age)
## factor to character to integer
escapes$Age.in.months <- as.integer(escapes$Age.in.months)
## plotting a histogram
age <- ggplot(data=escapes, aes(x=Age.in.months)) + labs(title =
"Participants by age in months", y = "Age in Months")
age <- age+geom_histogram(binwidth = 2)
age
```



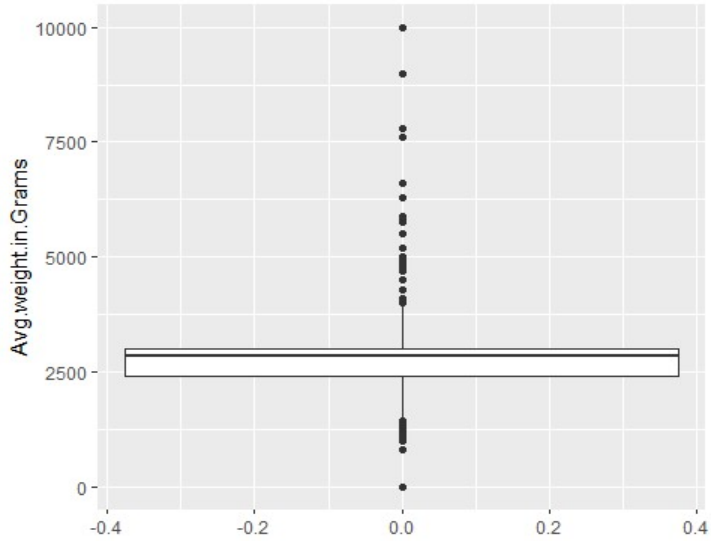

```
# Checking the quantiles of columns
quantile(escapes$Age.in.months)

##   0%  25%  50%  75% 100%
##    2   12   15   16   48
```

Average weight

```
escapes$Average.Weight <- str_remove_all(escapes$Average.Weight, " ")
escapes$Average.Weight <- str_remove_all(escapes$Average.Weight, "~")
escapes$Average.Weight[escapes$Average.Weight %in% c("9months","unknown",
"postsmolt", "6.5&12kg",  "15months(insw)")] <- "NA"
## Kgs to Grams
convertGrams <- function (svalue){
   if(grepl("kg", svalue)){ # find the position of the k (if any)
      kg <- str_locate(svalue,"kg")[1,1]
      num1 = as.numeric(substr(svalue, 0,kg-1)) # get 1st number
```

```
      return(num1*1000)
   } else {return(as.character(svalue))} # no range specified
}
escapes$Average.Weight <- sapply((escapes$Average.Weight), convertGrams)

## Warning in FUN(X[[i]], ...): NAs introduced by coercion

## Warning in FUN(X[[i]], ...): NAs introduced by coercion

## Warning in FUN(X[[i]], ...): NAs introduced by coercion

## Warning in FUN(X[[i]], ...): NAs introduced by coercion

## Warning in FUN(X[[i]], ...): NAs introduced by coercion

## Warning in FUN(X[[i]], ...): NAs introduced by coercion

## Warning in FUN(X[[i]], ...): NAs introduced by coercion

## Warning in FUN(X[[i]], ...): NAs introduced by coercion

escapes$Average.Weight <- sapply((escapes$Average.Weight), findMean)## Ranges
replacing by mean

## Warning in FUN(X[[i]], ...): NAs introduced by coercion

## Warning in FUN(X[[i]], ...): NAs introduced by coercion

## Warning in FUN(X[[i]], ...): NAs introduced by coercion

## Warning in FUN(X[[i]], ...): NAs introduced by coercion

## Warning in FUN(X[[i]], ...): NAs introduced by coercion

## Warning in FUN(X[[i]], ...): NAs introduced by coercion

## Warning in FUN(X[[i]], ...): NAs introduced by coercion

## Warning in FUN(X[[i]], ...): NAs introduced by coercion

## Warning in FUN(X[[i]], ...): NAs introduced by coercion

## Warning in FUN(X[[i]], ...): NAs introduced by coercion

## Warning in FUN(X[[i]], ...): NAs introduced by coercion

escapes$Average.Weight <- as.numeric(escapes$Average.Weight)## Character to
numeric
```

```
## Warning: NAs introduced by coercion

escapes$Average.Weight <- round(impute((escapes$Average.Weight), mean))## 
Imputing mean for NA's
colnames(escapes)[colnames(escapes) == "Average.Weight"] <- 
"Avg.weight.in.Grams" ## Renaming column name
escapes$Average.weight<- drop(escapes$Average.Weight)
escapes$Avg.weight.in.Grams <- as.integer(escapes$Avg.weight.in.Grams)## 
Numeric to integer
p <- ggplot(data=escapes, aes(y=Avg.weight.in.Grams))## plotting a boxplot
p <- p+geom_boxplot()
p
```




```
IQR(escapes$Avg.weight.in.Grams)## Inter quartile range

## [1] 600

range(escapes$Avg.weight.in.Grams) ## Range

## [1]     1 10000

var(escapes$Avg.weight.in.Grams) ## Checking distribution spread (variance)

## [1] 1496411

sd(escapes$Avg.weight.in.Grams) ## standard deviation

## [1] 1223.279
```

Escape.Start.Date

```
escapes$Escape.Start.Date <- dmy(escapes$Escape.Start.Date) ## character to 
date datatype
escapes$yearMonth <- format(escapes$Escape.Start.Date, "%Y-%m")## Splitting 
Date column into three seperate columns
escapes$year <- format(escapes$Escape.Start.Date, "%Y")
```

```
escapes$month <- format(escapes$Escape.Start.Date, "%m")
escapes$month <- str_remove(escapes$month, "^0+") ## Removing "0" from months column
escapes$yearMonth <- paste(escapes$year, escapes$month, sep = "-")
```

Discarding unimportant features/important

```
## Dropping Escape.Start.Date
escapes$Escape.Start.Date <- NULL ## This has splitted into separate columns
escapes$Escape.Water.Type <- NULL ## No useful information with 3 levels
escapes$Escape.Start.Time <- NULL ## This may be preprocessed but again assuming dates are suffice
escapes$Escape.End.Time <- NULL ## No useful information for the specific task
escapes$Escape.Grid.Reference <- NULL ## Not important
escapes$Stage <- NULL ## Unimportant
escapes$Initial.Date.of.Escape <- NULL ## Unimportant
escapes$Initial.Number.Escaped <- NULL ## Unimportant
escapes$Initial.Escape.Reason <- NULL ## Unimportant as we have final escape reason
escapes$Final.Date.of.Escape <- NULL ## Unimportant as we have escape start date preprocessed
escapes$Date.Registered <- NULL ## Unimportant
escapes$National.Grid.Reference <- NULL ## Unimportant
escapes$Local.Authority <- NULL ## Unimportant
escapes$Site.Address.1 <- NULL ## Unimportant
escapes$Site.Address.2 <- NULL ## Unimportant
escapes$Site.Address.3 <- NULL ## Unimportant
escapes$Site.Post.Code <- NULL ## Unimportant
escapes$Site.Contact.Number <- NULL ## Unimportant
escapes$Aquaculture.Type <- NULL ## Unimportant only one level
escapes$Easting <- NULL ## Unimportant
escapes$Northing <- NULL ## Unimportant
escapes$MS.Management.Area <- NULL ## Unimportant
escapes$Region <- NULL ## Unimportant
escapes$Operator <- NULL ## Unimportant as we have operator at time of escapes
escapes$Species <- NULL ## Data is inconsistent as one species is repeated a lot of time in every instance also we have Escaped species
```

Escape ID

```
escapes$Escape.ID <- as.factor(escapes$Escape.ID)## Changing the data type for Escape Id from integer to factor
str(escapes$Escape.ID) ##changed from integer to factor and it has 357 levels with no duplicates

##  Factor w/ 357 levels "2000001","2000023",..: 259 162 53 354 180 230 177 1 142 40 ...
```

Operator at time of escape

```
escapes$Operator.at.Time.of.Escape[escapes$Operator.at.Time.of.Escape ==
"aquascot sea farms ltd."] <- "aquascot sea farms ltd"
escapes$Operator.at.Time.of.Escape[escapes$Operator.at.Time.of.Escape ==
"ardvar salmon"] <- "ardvar salmon ltd"
escapes$Operator.at.Time.of.Escape[escapes$Operator.at.Time.of.Escape ==
"ardvar salmon ltd."] <- "ardvar salmon ltd"
escapes$Operator.at.Time.of.Escape[escapes$Operator.at.Time.of.Escape ==
"caledonian trout co" ] <- "caledonian trout company"
escapes$Operator.at.Time.of.Escape[escapes$Operator.at.Time.of.Escape ==
"cloan hatcheries"] <- "cloan hatcheries ltd."
escapes$Operator.at.Time.of.Escape[escapes$Operator.at.Time.of.Escape ==
"drummond fish farms"] <- "drummond fish farms ltd."
escapes$Operator.at.Time.of.Escape[escapes$Operator.at.Time.of.Escape ==
"kames fish farming ltd."] <- "kames fish farming ltd"
escapes$Operator.at.Time.of.Escape[escapes$Operator.at.Time.of.Escape ==
"kames marine fish farming" ] <- "kames fish farming ltd"
escapes$Operator.at.Time.of.Escape[escapes$Operator.at.Time.of.Escape ==
"landcatch ltd" ] <- "landcatch natural selection ltd"
escapes$Operator.at.Time.of.Escape[escapes$Operator.at.Time.of.Escape ==
"mainland salmon"] <- "mainland salmon ltd"
escapes$Operator.at.Time.of.Escape[escapes$Operator.at.Time.of.Escape ==
"marine harvest (scotland) ltd."] <- "marine harvest (scotland) ltd"
escapes$Operator.at.Time.of.Escape[escapes$Operator.at.Time.of.Escape ==
"murray seafood ltd."] <- "murray seafoods ltd"
escapes$Operator.at.Time.of.Escape[escapes$Operator.at.Time.of.Escape ==
"murray seafoods ltd."] <- "murray seafoods ltd"
escapes$Operator.at.Time.of.Escape[escapes$Operator.at.Time.of.Escape ==
"pan fish scotland ltd."] <- "pan fish scotland ltd"
escapes$Operator.at.Time.of.Escape[escapes$Operator.at.Time.of.Escape ==
"papil salmon farms ltd."] <-  "skelda salmon farms ltd"
escapes$Operator.at.Time.of.Escape[escapes$Operator.at.Time.of.Escape ==
"skelda salmon" ] <- "papil salmon farm ltd"
escapes$Operator.at.Time.of.Escape[escapes$Operator.at.Time.of.Escape ==
"stolt sea farm ltd."] <- "stolt seafarm ltd."
escapes$Operator.at.Time.of.Escape[escapes$Operator.at.Time.of.Escape ==
"stolt sea farms ltd." ] <- "stolt seafarm ltd."
escapes$Operator.at.Time.of.Escape[escapes$Operator.at.Time.of.Escape ==
"torhouse trout"] <- "torhouse trout ltd"
levels(escapes$Operator.at.Time.of.Escape) <-
droplevels(escapes$Operator.at.Time.of.Escape) ## Dropping all noisy levels
that has been preprocessed
```

Final number escaped

```
escapes$Final.Number.Escaped <- str_remove_all(escapes$Final.Number.Escaped,
" ")
escapes$Final.Number.Escaped <- str_remove_all(escapes$Final.Number.Escaped,
"~")
escapes$Final.Number.Escaped <- str_remove_all(escapes$Final.Number.Escaped,
```

```
",")
escapes$Final.Number.Escaped[escapes$Final.Number.Escaped %in% c("unknown",
"ca.150", "0(160000dead)", "0(13dead)" , ">500<1050" , "unnown", "none" ,
"nolosssuspected" , "20(estimate)", "zero" )] <- NA
escapes$Final.Number.Escaped <- sapply((escapes$Final.Number.Escaped),
findMean) ## Imputing mean values for ranges

## Warning in FUN(X[[i]], ...): NAs introduced by coercion

escapes$Final.Number.Escaped <- as.numeric(escapes$Final.Number.Escaped)
escapes$Final.Number.Escaped <- round(impute((escapes$Final.Number.Escaped),
mean))
escapes$Final.Number.Escaped <- as.integer(escapes$Final.Number.Escaped)##
Imputing mean for NA's
escapes$Final.Number.Escaped<- drop(escapes$Final.Number.Escaped)
```

Final number recovered

```
escapes$Final.Number.Recovered <-
str_remove_all(escapes$Final.Number.Recovered, " ")
escapes$Final.Number.Recovered <-
str_remove_all(escapes$Final.Number.Recovered, "~")
escapes$Final.Number.Recovered <-
str_remove_all(escapes$Final.Number.Recovered, ",")
escapes$Final.Number.Recovered[escapes$Final.Number.Recovered %in% c("<NA>",
" n/a", "none - n/a no loss", "none" , "+" , "15 live ", "unkonwn, local
anglers catching fish" ,  "zero" )] <- NA
escapes$Final.Number.Recovered <- as.numeric(escapes$Final.Number.Recovered)

## Warning: NAs introduced by coercion

escapes$Final.Number.Recovered<-
round(impute((escapes$Final.Number.Recovered), mean)) ## Imputing mean for
NA's
escapes$Final.Number.Recovered <- as.integer(escapes$Final.Number.Recovered)
escapes$Final.Number.Recovered <- drop(escapes$Final.Number.Recovered)
```

Final escape reason

```
escapes$Final.Escape.Reason <- as.factor(escapes$Final.Escape.Reason) ##
factor from character
```

Health surveillance

```
escapes$Health.Surveillance[escapes$Health.Surveillance == "not applicable"]
<- "NA" ## NA's and "Not applicable"

## Warning in `[<-.factor`(`*tmp*`, escapes$Health.Surveillance == "not
## applicable", : invalid factor level, NA generated
```

Data preparation and exploring Analysis dataset:

This is about escape incidents analysis that has been carried out to monitor the incidents. looking into it the data set has 9 columns/attributes two integer for year and month,

numeric for analysis c2, c3, c4, c5, c6, c7 and Site name is factor datatype. The data set looks clean and no NA's or missing values. Overall looking into dimensions data set has 351 rows and 9 columns

```
summary(analysis)

##       year          month                  Site.Name         c2        
##  Min.   :1998   Min.   : 1.000   Balta Isle     :  8   Min.   :0.2800  
##  1st Qu.:2005   1st Qu.: 3.000   Eilean Grianain:  8   1st Qu.:0.8775  
##  Median :2008   Median : 6.000   Loch Earn      :  8   Median :1.4070  
##  Mean   :2009   Mean   : 6.405   Loch Tay       :  8   Mean   :1.4082  
##  3rd Qu.:2015   3rd Qu.:10.000   Loch Lochy     :  6   3rd Qu.:1.9245  
##  Max.   :2020   Max.   :12.000   Taranaish      :  6   Max.   :2.6990  
##                                  (Other)        :307                   
##        c3               c4               c5                c6         
##  Min.   :0.0000   Min.   :0.1943   Min.   :0.01528   Min.   :0.03842  
##  1st Qu.:0.1440   1st Qu.:0.6376   1st Qu.:0.01738   1st Qu.:0.07764  
##  Median :0.3220   Median :1.0420   Median :0.01804   Median :0.09299  
##  Mean   :0.3384   Mean   :1.0600   Mean   :0.01805   Mean   :0.09336  
##  3rd Qu.:0.5175   3rd Qu.:1.4603   3rd Qu.:0.01877   3rd Qu.:0.10709  
##  Max.   :0.9600   Max.   :2.2184   Max.   :0.02062   Max.   :0.16728  
##                                                                        
##        c7         
##  Min.   :0.02807  
##  1st Qu.:0.08614  
##  Median :0.15394  
##  Mean   :0.15759  
##  3rd Qu.:0.21460  
##  Max.   :0.38359  
##                   

dim(analysis)

## [1] 351   9
```

TASK 2

Integrating both data sets escapes.csv and analysis.csv. We will merge by yearMonth column using retaining only rows where there are matching records in both data sets. That is the yearMonth appears in both dataset that has been combined together from two separate columns. Little processing is required where the site.Name has been converted to lower case to match the escapes data set. Year and month merged into one column as yearMonth using paste function separated by "-". Using merge function both data sets is merged where x is assigned to data sets "escapes" and y is assigned to "analysis". Merge function uses "by.x" and "by.y" which is assigned to common columns "yearMonth" and "Site.Name" which finds all the matching records and doesn't include those records where there is no match. This merged dataset will be saved in "escapesPlus.csv" file using write function.

```
analysis$Site.Name <- as.factor(tolower(analysis$Site.Name)) ## Lower casing
analysis$yearMonth <- paste(analysis$year, analysis$month, sep = "-") ##
Merging year and month
##analysis$month <- as.character(month(analysis$month))
 escapesPlus <- merge(x = escapes, y = analysis,
                      by.x =  c("yearMonth", "Site.Name"),
                      by.y =  c("yearMonth", "Site.Name"),
                      all.x = FALSE, all.y = FALSE)
## Writing file to csv
## write <- write.csv(escapesPlus, file = "escapesPlus.csv", row.names =
FALSE)
```

TASK 3

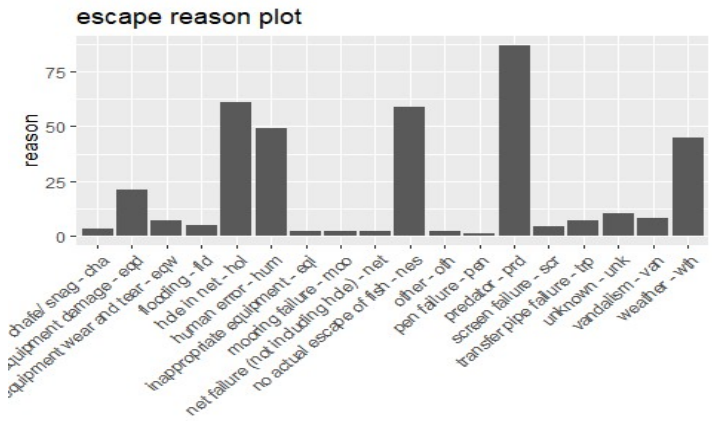
- Exploratory data analysis and preparation of escapesPlus. Has 375 rows and 24 columns. We have same NA's in nominal attributes "Final escape reason" and "Health surveillance" overall we have 21 NA's. yearMonth is the column that we used to merge the data sets is character datatype. We have 8 factor, 10 numeric and 3 character variables. In the below tasks we are undertaking data preparation tasks and data cleaning with some analysis showing specifically relation of focused variables with other.
- Numeric attributes 5 number summary statistics ageInMonths is normally distributed as mean and median is close as 15 months of escaped species age. WeightGrams is normally distributed too with mean and median quite close with 2830 grams as median and 2807 as mean. numberofEscaped and "numberofRecovered" mean and median is not close to each other representing skewed distribution. escapeReason and siteID are distributed normally while producingInLast3yrs is imbalanced with "yes" class highly distributed than "no". waterType has outlier "freshwater and seawater" and is imbalanced. healthSurveillance is also not normally distributed while the analysis attributes, c2,c3,c4,c5,c6 and c7 mean and median is close so they are normally distributed.
- Missing values has been imputed with mode values for categorical features. and duplicate columns as such "year", "month" and "siteID" has been discarded. Also column name has been renamed to interpret them much easier way.
- Univariate analysis of "escape reason plot" ggplot has been plotted to see the distribution of the levels where we can see "predator-prd" is contributing majority among escaped species, "no actual reason" and "hole in the net" is impacting lesser than predator-prd.
- Plot "Distribution of escaped species age and water type" shows species "halibut" of age between 24 and 40 months who live in sea water is highly distributed in the attribute. "Rainbow trout and atlantic salmon has a lot of outliers.
- Distribution of escaped species on sites has been plotted as well as number of escaped species recovered with average weight has also been plotted, that gives an interesting information in regard to how many species recovered on what sites "loch greshornish" has recovered maximum amount of escaped species.

```
## Reshaping data and renaming columns
colnames(escapesPlus)[colnames(escapesPlus) == "Operator.at.Time.of.Escape"]
<- "operator"
colnames(escapesPlus)[colnames(escapesPlus) == "Site.Name"] <- "siteName"
colnames(escapesPlus)[colnames(escapesPlus) == "Escape.ID"] <- "escapeID"
colnames(escapesPlus)[colnames(escapesPlus) == "Escaped.Species"] <-
"escapedSpecies"
colnames(escapesPlus)[colnames(escapesPlus) == "Age.in.months"] <-
"ageInMonths"
colnames(escapesPlus)[colnames(escapesPlus) == "Avg.weight.in.Grams"] <-
"weightGrams"
colnames(escapesPlus)[colnames(escapesPlus) == "Final.Number.Escaped"] <-
"numberOfEscaped"
colnames(escapesPlus)[colnames(escapesPlus) == "Final.Number.Recovered"] <-
"numberofRecovered"
colnames(escapesPlus)[colnames(escapesPlus) == "Final.Escape.Reason"] <-
"escapeReason"
colnames(escapesPlus)[colnames(escapesPlus) == "Marine.Scotland.Site.ID"] <-
"siteID"
colnames(escapesPlus)[colnames(escapesPlus) == "Producing.in.Last.3.Years"]
<- "producingInLast3Yrs"
colnames(escapesPlus)[colnames(escapesPlus) == "Water.Type"] <- "waterType"
colnames(escapesPlus)[colnames(escapesPlus) == "Health.Surveillance"] <-
"healthSurveillance"
colnames(escapesPlus)[colnames(escapesPlus) == "year.x"] <- "year"
colnames(escapesPlus)[colnames(escapesPlus) == "month.x"] <- "month"
escapesPlus$year.y <- NULL ## Duplicate column
escapesPlus$month.y <- NULL ## Duplicate column
escapesPlus$siteID <- NULL ## Site name is suffice
dim(escapesPlus)

## [1] 375  21

summary(escapesPlus) ## An overview to see what preparation required
including missing values

##   yearMonth                    siteName      escapeID
##  Length:375         eilean grianain: 12   2000483:  3
##  Class :character   balta isle     : 10   2000484:  3
##  Mode  :character   corlarach      : 10   2000485:  3
##                     loch tay       : 10   2000040:  2
##                     loch earn      :  8   2000041:  2
##                     taranaish      :  8   2000073:  2
##                     (Other)        :317   (Other):360
##                          operator                       escapedSpecies
##  david m brien                : 60   atlantic salmon          :294
##  the scottish salmon company  : 48   brown trout and sea trout:  1
##  kames fish farming ltd       : 35   cod                      :  1
##  marine harvest (scotland) ltd: 28   halibut                  :  2
##  ferramus (ss)                : 24   lumpsucker               :  1
```

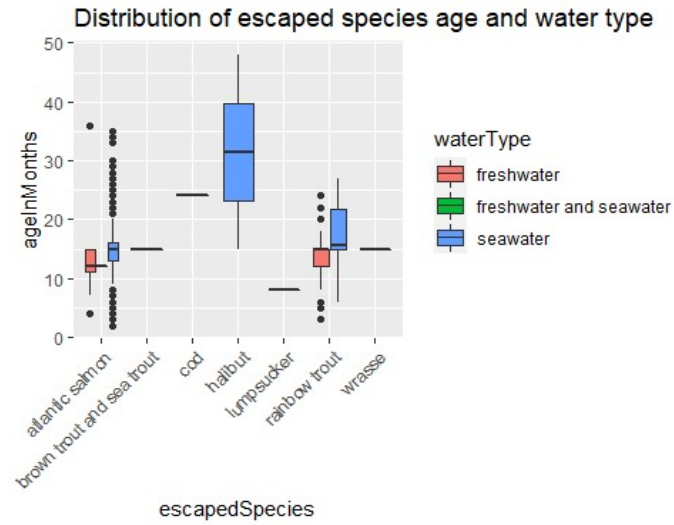
```
##  hjaltland seafarms ltd       : 24   rainbow trout            : 74
##  (Other)                      :156   wrasse                   :  2
##   ageInMonths     weightGrams    numberOfEscaped  numberofRecovered
##  Min.   : 2.00   Min.   :    1   Min.   :     0   Min.   :    0.0
##  1st Qu.:12.00   1st Qu.: 2370   1st Qu.:     0   1st Qu.:    0.0
##  Median :15.00   Median : 2830   Median :  1062   Median :    0.0
##  Mean   :15.02   Mean   : 2807   Mean   : 11108   Mean   :  197.2
##  3rd Qu.:16.00   3rd Qu.: 3000   3rd Qu.: 11376   3rd Qu.:  203.0
##  Max.   :48.00   Max.   :10000   Max.   :336470   Max.   :27453.0
##
##                          escapeReason producingInLast3Yrs
##  predator - prd                :82    no : 52
##  hole in net - hol             :61    yes:323
##  no actual escape of fish - nes:59
##  human error - hum             :49
##  weather - wth                 :45
##  (Other)                       :74
##  NA's                          : 5
##                    waterType        healthSurveillance     year
##  freshwater              : 85   high          : 35      Length:375
##  freshwater and seawater:  1   low           :123      Class :character
##  seawater                :289   medium        :201      Mode  :character
##                                 not applicable:  0
##                                 NA's          : 16
##
##
##     month                 c2               c3               c4
##  Length:375         Min.   :0.2800   Min.   :0.0000   Min.   :0.1943
##  Class :character   1st Qu.:0.8505   1st Qu.:0.1390   1st Qu.:0.6129
##  Mode  :character   Median :1.4070   Median :0.3220   Median :1.0389
##                     Mean   :1.4103   Mean   :0.3409   Mean   :1.0649
##                     3rd Qu.:1.9575   3rd Qu.:0.5330   3rd Qu.:1.4744
##                     Max.   :2.6990   Max.   :0.9600   Max.   :2.2184
##
##        c5                c6                c7
##  Min.   :0.01528   Min.   :0.03842   Min.   :0.02807
##  1st Qu.:0.01739   1st Qu.:0.07716   1st Qu.:0.08490
##  Median :0.01805   Median :0.09288   Median :0.15054
##  Mean   :0.01806   Mean   :0.09329   Mean   :0.15738
##  3rd Qu.:0.01876   3rd Qu.:0.10735   3rd Qu.:0.21621
##  Max.   :0.02062   Max.   :0.16728   Max.   :0.38359
##
```

```
## Imputing mode for NA's
escapesPlus$escapeReason <- impute((escapesPlus$escapeReason), mode)
escapesPlus$healthSurveillance <- impute((escapesPlus$healthSurveillance),
mode)
## bar chart with rotated labels for escape reason
ggplot(escapesPlus, aes(x = escapeReason)) +
  geom_bar() +
```

```
  labs(x = "",
       y = "reason",
       title = "escape reason plot") +
  theme(axis.text.x = element_text(angle = 45,
                                   hjust = 1))
```

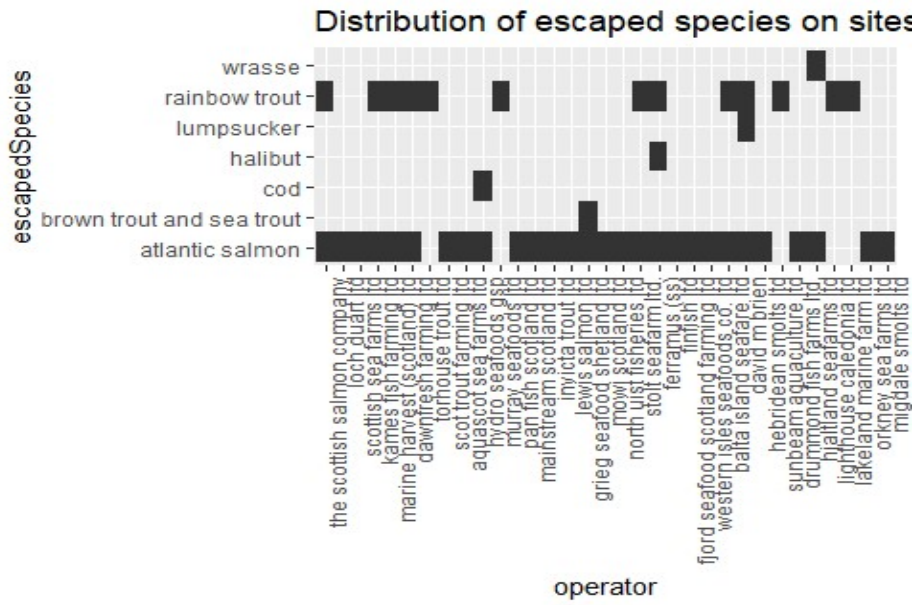



```
# boxplot for ageinmonths escapedspecies and watertype
ggplot(escapesPlus) +
  aes(x = ageInMonths, y = escapedSpecies, fill = waterType) +
  geom_boxplot() +
   labs( title = "Distribution of escaped species age and water type") +
  scale_fill_hue(direction = 1) +
  coord_flip() +
  theme_gray()+
  theme(axis.text.x = element_text(angle = 45,
                                   hjust = 1))
```



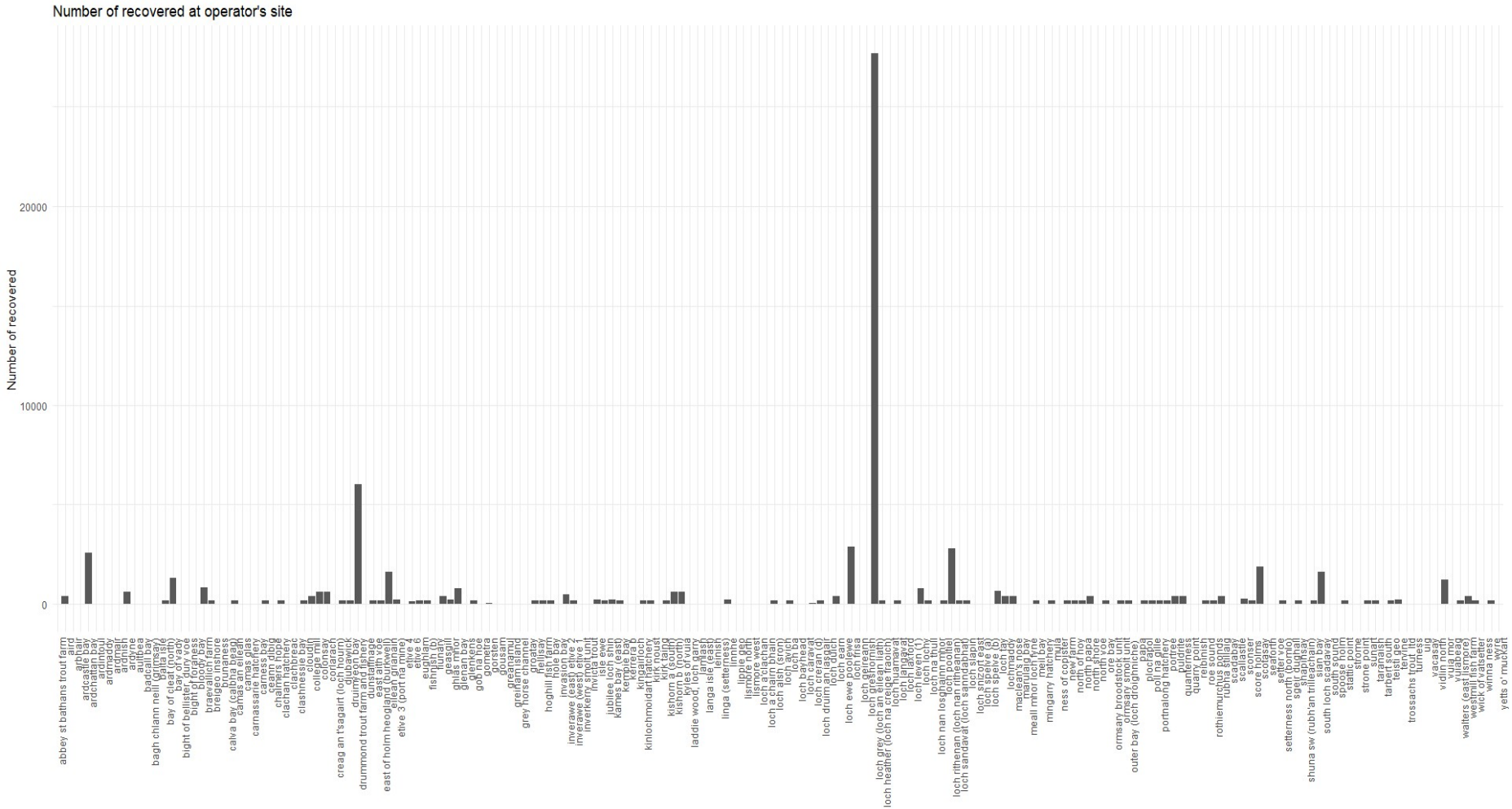

```
# plot species weight with final recovered
ggplot(escapesPlus) +
  aes(x = operator, y = escapedSpecies) +
```

```
   labs(title ="Distribution of escaped species on sites") +
  geom_tile(size = 1.5) +
  theme_gray() +
  theme(axis.text.x = element_text(angle = 90,
                                   hjust = 1))
```




```
ggplot(escapesPlus) +
 aes(x = siteName, weight = numberofRecovered) +
 geom_bar(fill = "#FF69B4") +
 theme_minimal() +
   labs( x = "", y = "Number of recovered",title = "Number of recovered at
operator's site") +
  geom_bar() +
  theme(axis.text.x = element_text(angle = 90,
                                   hjust = 1))
```

Number of recovered species at operator's site:

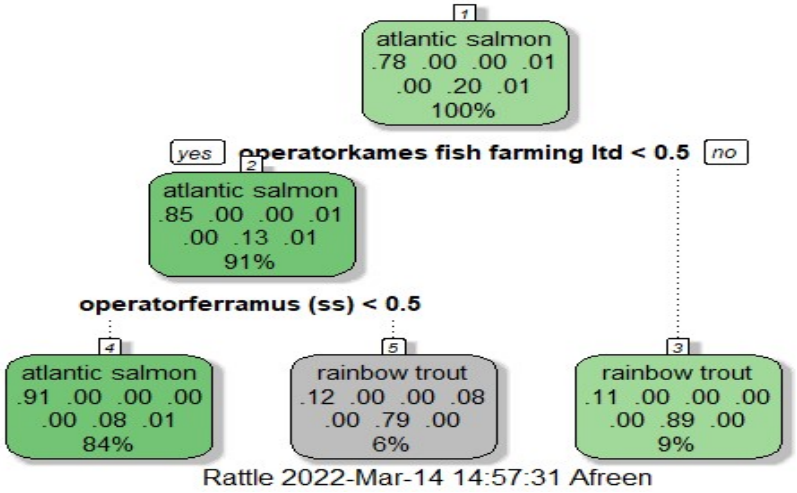


TASK 4

- Looking at our data set and business requirement the Aquaculture's fishes are kept in cages and incidents referring to escapes has been monitored for continuous improvement in order to avoid these escapes. Overall we want to predict feature escapedSpecies using the rest of the data in escapesPlus.
- In this task we will run an experiment using rpart and random forest to predict escapedSpecies. Setting train controls "out of bag" and "cross validation" In order to compare tree-building algorithms.
- Two train controls has been defined one uses out of bag error and second is repeated 10 fold cross validation. Rpart:- The best accuracy was 0.8863145 with a corresponding kappa of 0.6366269 (moderate agreement). The confusion matrix shows that 74.9% of the instances corresponded to class Atlantic salmon correctly predicted and 13.6% of rainbow trout correctly predicted.Also there has been misclassification of atlantic salmon missclassified as 0.3% of codsea trout, lumpsucker and rainbow trout. While rainbow trout has been correctly classified but also misclassified as Atlantic salmon for 3.5%
- Random forest:- The accuracy is a little higher than for rpart with 93.2% (compared to 88%) with a kappa value showing moderate agreement (0.8092216). The distribution of errors is quite different, with proportionally more errors where class "atlantic salmon" instances were misclassified but less where class "rainbow trout" instances were misclassified.
- Variable Importance:- There are 3 features which has quite a lot of importance and impact including operatorkames fish farming ltd from attribute "operator" is really important and waterTypeseawater and operatorferramus (ss) from are lesser important than operatorkames. Other 17 features are less important or have less impact.

```
trainControl <- trainControl(method = "oob") ## Train control out of bag
trainControl2 <- trainControl(method="repeatedcv", number=10, repeats=1) ##
Repeated cv
## rpart
set.seed(123)
rpart.escapes <- train(escapedSpecies~.,
 data = escapesPlus,
 method = "rpart",
 metric = "Accuracy",
 trControl = trainControl2)
print(rpart.escapes)

## CART
##
## 375 samples
##  20 predictor
##   7 classes: 'atlantic salmon', 'brown trout and sea trout', 'cod',
'halibut', 'lumpsucker', 'rainbow trout', 'wrasse'
##
```

```
## No pre-processing
## Resampling: Cross-Validated (10 fold, repeated 1 times)
## Summary of sample sizes: 336, 338, 338, 338, 338, 339, ...
## Resampling results across tuning parameters:
##
##   cp          Accuracy   Kappa
##   0.02469136  0.8863145  0.6366269
##   0.19753086  0.8544225  0.4484499
##   0.33333333  0.8027669  0.1415602
##
## Accuracy was used to select the optimal model using the largest value.
## The final value used for the model was cp = 0.02469136.

confusionMatrix.train(rpart.escapes) ## Confusion matrix

## Cross-Validated (10 fold, repeated 1 times) Confusion Matrix
##
## (entries are percentual average cell counts across resamples)
##
##                            Reference
## Prediction                  atlantic salmon brown trout and sea trout  cod
##   atlantic salmon                     74.9                       0.3  0.3
##   brown trout and sea trout            0.0                       0.0  0.0
##   cod                                  0.0                       0.0  0.0
##   halibut                              0.0                       0.0  0.0
##   lumpsucker                           0.0                       0.0  0.0
##   rainbow trout                        3.5                       0.0  0.0
##   wrasse                               0.0                       0.0  0.0
##                            Reference
## Prediction                  halibut lumpsucker rainbow trout wrasse
##   atlantic salmon               0.0        0.3           6.1    0.5
##   brown trout and sea trout     0.0        0.0           0.0    0.0
##   cod                           0.0        0.0           0.0    0.0
##   halibut                       0.0        0.0           0.0    0.0
##   lumpsucker                    0.0        0.0           0.0    0.0
##   rainbow trout                 0.5        0.0          13.6    0.0
##   wrasse                        0.0        0.0           0.0    0.0
##
##  Accuracy (average) : 0.8853

fancyRpartPlot(rpart.escapes$finalModel) ## plotting rplot
```




```
## Random forest
set.seed(123)
rf.escapes <- train(escapedSpecies~.,
 data = escapesPlus,
 method = "rf",
 metric = "Accuracy",
 ntree = 50,
 trControl = trainControl2)

## Warning: model fit failed for Fold03.Rep1: mtry=  2 Error in
randomForest.default(x, y, mtry = min(param$mtry, ncol(x)), ...) :
##   Can't have empty classes in y.

## Warning: model fit failed for Fold03.Rep1: mtry= 41 Error in
randomForest.default(x, y, mtry = min(param$mtry, ncol(x)), ...) :
##   Can't have empty classes in y.

print(rf.escapes)

## Random Forest
##
## 375 samples
##  20 predictor
##   7 classes: 'atlantic salmon', 'brown trout and sea trout', 'cod',
'halibut', 'lumpsucker', 'rainbow trout', 'wrasse'
##
## No pre-processing
## Resampling: Cross-Validated (10 fold, repeated 1 times)
## Summary of sample sizes: 336, 338, 338, 338, 338, 339, ...
## Resampling results across tuning parameters:
##
##   mtry  Accuracy   Kappa
##     2   0.7877108  0.0000000
##    41   0.8981751  0.6495223
##   846   0.9281787  0.7727189
```

```
##
## Accuracy was used to select the optimal model using the largest value.
## The final value used for the model was mtry = 846.

confusionMatrix.train(rf.escapes) #Confusion matrix

## Cross-Validated (10 fold, repeated 1 times) Confusion Matrix
##
## (entries are percentual average cell counts across resamples)
##
##                            Reference
## Prediction                  atlantic salmon brown trout and sea trout  cod
##   atlantic salmon                     76.4                       0.0  0.0
##   brown trout and sea trout            0.0                       0.0  0.0
##   cod                                  0.0                       0.0  0.0
##   halibut                              0.0                       0.0  0.0
##   lumpsucker                           0.0                       0.0  0.0
##   rainbow trout                        2.3                       0.0  0.0
##   wrasse                               0.0                       0.0  0.0
##                            Reference
## Prediction                  halibut lumpsucker rainbow trout wrasse
##   atlantic salmon               0.0        0.0           4.6    0.0
##   brown trout and sea trout     0.0        0.0           0.0    0.0
##   cod                           0.0        0.0           0.0    0.0
##   halibut                       0.0        0.0           0.0    0.0
##   lumpsucker                    0.0        0.0           0.0    0.0
##   rainbow trout                 0.4        0.0          15.6    0.0
##   wrasse                        0.0        0.0           0.0    0.8
##
##  Accuracy (average) : 0.9278

## Variable importance
varImp(rf.escapes)

## rf variable importance
##
##   only 20 most important variables shown (out of 847)
##
##                                  Overall
## operatorkames fish farming ltd   100.000
## operatorferramus (ss)             61.851
## waterTypeseawater                 59.120
## operatormurray seafoods ltd       12.015
## numberOfEscaped                    9.869
## escapeID2000453                    8.706
## c5                                 7.779
## escapeReasonflooding - fld         6.400
## escapeReasonvandalism - van        6.262
## ageInMonths                        5.884
## c3                                 4.786
## c7                                 4.639
```

```
## producingInLast3Yrsyes              4.458
## operatorlakeland marine farm ltd    4.376
## c6                                  4.216
## escapeID2000451                     4.178
## c4                                  3.814
## yearMonth2007-3                     3.658
## yearMonth2017-3                     2.893
## siteNamemeil bay                    2.835
```

References

Gavin Simpsons - Split date data (m/d/y) into 3 separate columns, 2010. *Stack over flow.* [Online]
Available at: https://stackoverflow.com/questions/4078502/split-date-data-m-d-y-into-3-separate-columns
[Accessed 10 03 2022].

Ines Arana - Lab 1,2,3,4,5 - RGU, 2022. *Exploratory data analysis, Data preparation and Random forest.* Aberdeen: RGU.

MASL - Stack over flow, 2015. *Stack over flow.* [Online]
Available at: https://datascience.stackexchange.com/posts/8924/timeline
[Accessed 10 03 2022].