MATH2349 Semester 1, 2019

Code

Assignment 3

Abdul Gazi 3705448, Akhil Puri 3774583 Lily Xia 3603966

Required packages

Hide

```
library(readr)
library(readxl)
library(foreign)
library(rvest)
library(dplyr)
library(tidyr)
library(stringr)
library(lubridate)
library(car)
library(MASS)
library(ggplot2)
library(knitr)
library(forecast)
```

```
Registered S3 method overwritten by 'xts':
  method     from
  as.zoo.xts zoo
Registered S3 method overwritten by 'quantmod':
  method            from
  as.zoo.data.frame zoo
Registered S3 methods overwritten by 'forecast':
  method             from
  fitted.fracdiff    fracdiff
  residuals.fracdiff fracdiff
This is forecast 8.7
  Stackoverflow is a great place to get help on R issues:
  http://stackoverflow.com/tags/forecasting+r.
```

Data

Plastic debris has been increasingly found contaminating oceans worldwide, posing a threat to both marine life and the surrounding animals that rely on them. This report details the preprocessing of 3 sets of data. In this report we are merging three data sets from surveys done in Australian waters. Vessel data contains the vessal used to collect the plastic, with information about who did it and when. Replicates contains information about whether or not the net station was repeated, and survey data contains information about the type of plastic collected on this explorations. All data has been sourced from https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0080466#s2 (https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0080466#s2), distributed unter the terms of the Creative Commons Attribution License. This report details the combining, preoprocessing and validation of the aforementioned datasets.

Importing

Explanation: The three data sets described above have been read into r and beened named accordingly (Vessel, Replicates and Survey). Head() has been utilised after each read step to show the first 6 lines of these datasets.

Two steps have been take to merge the data. Initially, 'Vessel' and 'Replicates' were merged together by date and named 'tempcombined' as this temporarily combined dataset will soon be merged with Survey as well. 'Finalcombined' is the product of Survey and 'tempcombined' being merged together (on the common 'NetStation' variable' and refers to the final dataset with all three datasets combined. The first 6 rows of the new 'finalcombined' dataset has been shown using head.

Hide

```
Vessel<- read.csv(file="vessels.csv",sep = ",", header =TRUE, stringsAsFactors = FALS
E )
getwd()
```

```
[1] "/Users/lilyxia/Downloads"
```

Hide

```
head(Vessel)
```

	VesselTripName <chr>	CollectedBy <chr>	NetType <chr>	Date <chr>
1	Enterprise/Austral Fisheries	Julia Reisser	Manta/ 333/ 1 x 0.17	26/07/2012
2	Enterprise/Austral Fisheries	Julia Reisser	Manta/ 333/ 1 x 0.17	27/07/2012
3	Enterprise/Austral Fisheries	Julia Reisser	Manta/ 333/ 1 x 0.17	28/07/2012
4	Enterprise/Austral Fisheries	Julia Reisser	Manta/ 333/ 1 x 0.17	29/07/2012
5	Enterprise/Austral Fisheries	Julia Reisser	Manta/ 333/ 1 x 0.17	30/07/2012
6	Solander/Trip 5597	Julia Reisser	Manta/ 333/ 1 x 0.17	17/08/2012

6 rows

Hide

```
Replicates<- read.csv(file="replicates.csv",sep = ",", header =TRUE, stringsAsFactors
= FALSE)
head(Replicates)
```

	NetStation <int>	Replicate <int>	DateUTC <chr>	StartTime <chr>	EndTime <chr>	Duration <int>
1	1	1	10/06/2011	9:12	9:27	15
2	1	2	10/06/2011	9:32	9:47	15
3	1	3	10/06/2011	9:49	10:04	15
4	2	1	11/06/2011	5:51	6:01	10
5	2	2	11/06/2011	8:55	9:10	15
6	2	3	11/06/2011	9:11	9:26	15

6 rows

Hide

```
Survey<- read.csv(file="survey.csv",sep = ",", header =TRUE, stringsAsFactors = FALSE
)
head(Survey)
```

	NetStation <int>	Replicate <int>	HardPlastics <int>	SoftPlastics <int>	PlasticLines <int>	Styrofoam <int>	Pellets <int>	T
1	1	1	1	0	0	0	0	
2	1	2	1	0	0	0	0	
3	1	3	1	0	0	0	0	
4	2	1	1	0	0	0	0	
5	2	2	3	0	0	0	0	
6	2	3	2	0	1	0	0	

6 rows | 1-9 of 11 columns

Hide

```
tempcombined <- merge(Vessel , Replicates, by.x = "Date", by.y="DateUTC")
finalcombined <- merge(Survey, tempcombined, by = "NetStation")
head(finalcombined)
```

	NetStation <int>	Replicate.x <int>	HardPlastics <int>	SoftPlastics <int>	PlasticLines <int>	Styrofoam <int>	Pellets <int>
1	1	2	1	0	0	0	0
2	1	2	1	0	0	0	0
3	1	2	1	0	0	0	0
4	1	3	1	0	0	0	0
5	1	3	1	0	0	0	0
6	1	3	1	0	0	0	0

6 rows | 1-9 of 19 columns

Understand

Explanation: The attributes and interal structure of all three original datasets, as well as the combined dataset have been accessed through attributes() and str() respectively.

Net type has been extracted from the combined dataset andhas been converted into a factor. As there are two different types of nets (Neuston & Manta), they have been categorised.

Hide

```
attributes(Vessel)
```

```
$names
[1] "VesselTripName" "CollectedBy"    "NetType"        "Date"

$class
[1] "data.frame"

$row.names
 [1]  1  2  3  4  5  6  7  8  9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27
28
```

Hide

```
str(Vessel)
```

```
'data.frame':   28 obs. of  4 variables:
 $ VesselTripName: chr  "Enterprise/Austral Fisheries" "Enterprise/Austral Fisheries"
"Enterprise/Austral Fisheries" "Enterprise/Austral Fisheries" ...
 $ CollectedBy   : chr  "Julia Reisser" "Julia Reisser" "Julia Reisser" "Julia Reisse
r" ...
 $ NetType       : chr  "Manta/ 333/ 1 x 0.17" "Manta/ 333/ 1 x 0.17" "Manta/ 333/ 1
x 0.17" "Manta/ 333/ 1 x 0.17" ...
 $ Date          : chr  "26/07/2012" "27/07/2012" "28/07/2012" "29/07/2012" ...
```

Hide

```
attributes(Replicates)
```

```
$names
[1] "NetStation" "Replicate"  "DateUTC"    "StartTime"  "EndTime"    "Duration"

$class
[1] "data.frame"

$row.names
  [1]   1   2   3   4   5   6   7   8   9  10  11  12  13  14  15  16  17  18  19  20
21  22  23  24  25  26  27  28  29  30  31  32  33  34  35
 [36]  36  37  38  39  40  41  42  43  44  45  46  47  48  49  50  51  52  53  54  55
56  57  58  59  60  61  62  63  64  65  66  67  68  69  70
 [71]  71  72  73  74  75  76  77  78  79  80  81  82  83  84  85  86  87  88  89  90
91  92  93  94  95  96  97  98  99 100 101 102 103 104 105
[106] 106 107 108 109 110 111 112 113 114 115 116 117 118 119 120 121 122 123 124 125
126 127 128 129 130 131 132 133 134 135 136 137 138 139 140
[141] 141 142 143 144 145 146 147 148 149 150 151 152 153 154 155 156 157 158 159 160
161 162 163 164 165 166 167 168 169 170 171
```

Hide

```
str(Replicates)
```

```
'data.frame':   171 obs. of  6 variables:
 $ NetStation: int  1 1 1 2 2 2 3 3 3 4 ...
 $ Replicate : int  1 2 3 1 2 3 1 2 3 1 ...
 $ DateUTC   : chr  "10/06/2011" "10/06/2011" "10/06/2011" "11/06/2011" ...
 $ StartTime : chr  "9:12" "9:32" "9:49" "5:51" ...
 $ EndTime   : chr  "9:27" "9:47" "10:04" "6:01" ...
 $ Duration  : int  15 15 15 10 15 15 15 15 15 15 ...
```

Hide

```
attributes(Survey)
```

```
$names
 [1] "NetStation"    "Replicate"     "HardPlastics"  "SoftPlastics"  "PlasticLines"
"Styrofoam"     "Pellets"       "TotalPlastics"
 [9] "DriftingWood"  "Pumice"        "Cs"

$class
[1] "data.frame"

$row.names
  [1]   1   2   3   4   5   6   7   8   9  10  11  12  13  14  15  16  17  18  19  20
21  22  23  24  25  26  27  28  29  30  31  32  33  34  35
 [36]  36  37  38  39  40  41  42  43  44  45  46  47  48  49  50  51  52  53  54  55
56  57  58  59  60  61  62  63  64  65  66  67  68  69  70
 [71]  71  72  73  74  75  76  77  78  79  80  81  82  83  84  85  86  87  88  89  90
91  92  93  94  95  96  97  98  99 100 101 102 103 104 105
[106] 106 107 108 109 110 111 112 113 114 115 116 117 118 119 120 121 122 123 124 125
126 127 128 129 130 131 132 133 134 135 136 137 138 139 140
[141] 141 142 143 144 145 146 147 148 149 150 151 152 153 154 155 156 157 158 159 160
161 162 163 164 165 166 167 168 169 170 171
```

Hide

```
str(Survey)
```

```
'data.frame':   171 obs. of  11 variables:
 $ NetStation   : int  1 1 1 2 2 2 3 3 3 4 ...
 $ Replicate    : int  1 2 3 1 2 3 1 2 3 1 ...
 $ HardPlastics : int  1 1 1 1 3 2 2 2 0 1 ...
 $ SoftPlastics : int  0 0 0 0 0 0 0 0 0 0 ...
 $ PlasticLines : int  0 0 0 0 0 1 0 0 0 0 ...
 $ Styrofoam    : int  0 0 0 0 0 0 0 0 0 0 ...
 $ Pellets      : int  0 0 0 0 0 0 0 0 0 0 ...
 $ TotalPlastics: int  1 1 1 1 3 3 2 2 0 1 ...
 $ DriftingWood : int  0 0 0 1 3 0 0 0 0 1 ...
 $ Pumice       : int  0 0 4 0 0 0 0 0 0 1 ...
 $ Cs           : num  654 692 567 950 2199 ...
```

Hide

```
attributes(finalcombined)
```

```
$names
 [1] "NetStation"     "Replicate.x"    "HardPlastics"   "SoftPlastics"   "PlasticLine
s"   "Styrofoam"      "Pellets"        "TotalPlastics"
 [9] "DriftingWood"   "Pumice"         "Cs"             "Date"           "VesselTripN
ame" "CollectedBy"    "NetType"        "Replicate.y"
[17] "StartTime"      "EndTime"        "Duration"

$row.names
  [1]   1   2   3   4   5   6   7   8   9  10  11  12  13  14  15  16  17  18  19  20
21  22  23  24  25  26  27  28  29  30  31  32  33  34  35
 [36]  36  37  38  39  40  41  42  43  44  45  46  47  48  49  50  51  52  53  54  55
56  57  58  59  60  61  62  63  64  65  66  67  68  69  70
 [71]  71  72  73  74  75  76  77  78  79  80  81  82  83  84  85  86  87  88  89  90
91  92  93  94  95  96  97  98  99 100 101 102 103 104 105
[106] 106 107 108 109 110 111 112 113 114 115 116 117 118 119 120 121 122 123 124 125
126 127 128 129 130 131 132 133 134 135 136 137 138 139 140
[141] 141 142 143 144 145 146 147 148 149 150 151 152 153 154 155 156 157 158 159 160
161 162 163 164 165 166 167 168 169 170 171 172 173 174 175
[176] 176 177 178 179 180 181 182 183 184 185 186 187 188 189 190 191 192 193 194 195
196 197 198 199 200 201 202 203 204 205 206 207 208 209 210
[211] 211 212 213 214 215 216 217 218 219 220 221 222 223 224 225 226 227 228 229 230
231 232 233 234 235 236 237 238 239 240 241 242 243 244 245
[246] 246 247 248 249 250 251 252 253 254 255 256 257 258 259 260 261 262 263 264 265
266 267 268 269 270 271 272 273 274 275 276 277 278 279 280
[281] 281 282 283 284 285 286 287 288

$class
[1] "data.frame"
```

Hide

```
str(finalcombined)
```

```
'data.frame':   288 obs. of  19 variables:
 $ NetStation    : int  1 1 1 1 1 1 1 1 1 2 ...
 $ Replicate.x   : int  2 2 2 3 3 3 1 1 1 1 ...
 $ HardPlastics  : int  1 1 1 1 1 1 1 1 1 1 ...
 $ SoftPlastics  : int  0 0 0 0 0 0 0 0 0 0 ...
 $ PlasticLines  : int  0 0 0 0 0 0 0 0 0 0 ...
 $ Styrofoam     : int  0 0 0 0 0 0 0 0 0 0 ...
 $ Pellets       : int  0 0 0 0 0 0 0 0 0 0 ...
 $ TotalPlastics : int  1 1 1 1 1 1 1 1 1 1 ...
 $ DriftingWood  : int  0 0 0 0 0 0 0 0 0 1 ...
 $ Pumice        : int  0 0 0 4 4 4 0 0 0 0 ...
 $ Cs            : num  692 692 692 567 567 ...
 $ Date          : chr  "10/06/2011" "10/06/2011" "10/06/2011" "10/06/2011" ...
 $ VesselTripName: chr  "Southern Surveyor/ss2011_t02" "Southern Surveyor/ss2011_t02"
"Southern Surveyor/ss2011_t02" "Southern Surveyor/ss2011_t02" ...
 $ CollectedBy   : chr  "Julia Reisser, Briony Hutton" "Julia Reisser, Briony Hutton"
"Julia Reisser, Briony Hutton" "Julia Reisser, Briony Hutton" ...
 $ NetType       : chr  "Neuston/ 335/ 1.2 x 0.6" "Neuston/ 335/ 1.2 x 0.6" "Neuston/
335/ 1.2 x 0.6" "Neuston/ 335/ 1.2 x 0.6" ...
 $ Replicate.y   : int  1 2 3 1 2 3 1 2 3 2 ...
 $ StartTime     : chr  "9:12" "9:32" "9:49" "9:12" ...
 $ EndTime       : chr  "9:27" "9:47" "10:04" "9:27" ...
 $ Duration      : int  15 15 15 15 15 15 15 15 15 15 ...
```

Hide

```
finalcombined$NetType <- factor(finalcombined$NetType ,labels= c("Neuston/ 335/ 1.2 x 0.6","Manta/ 333/ 1 x 0.17"))
```

Tidy & Manipulate Data I

Explanation: In order to check whether the data conforms to tidy data principles, glimpse has been used to see every column in the final combined data frame. Head and tail have been used to view the first and last 10 rows of the data.

Hide

```
glimpse(finalcombined)
```

```
Observations: 288
Variables: 19
$ NetStation      [3m[38;5;246m<int> [39m[23m 1, 1, 1, 1, 1, 1, 1, 1, 1, 2, 2, 2, 2, 2, 2, 2, 2, 3, 3, 3, 3, 3, 3, 3, 3, 3, 4, 4, 4, 4, 4, 4, 4, 4, 4, 12, 12, 12, 1…
$ Replicate.x     [3m[38;5;246m<int> [39m[23m 2, 2, 2, 3, 3, 3, 1, 1, 1, 1, 1, 1, 2, 2, 2, 3, 3, 3, 1, 1, 1, 3, 3, 3, 2, 2, 2, 3, 3, 3, 1, 1, 1, 2, 2, 2, 3, 3, 3, 1, 1…
$ HardPlastics    [3m[38;5;246m<int> [39m[23m 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 3, 3, 3, 2, 2, 2, 2, 2, 2, 0, 0, 0, 2, 2, 2, 2, 2, 2, 1, 1, 1, 0, 0, 0, 1, 1, 1, 0, 0…
$ SoftPlastics    [3m[38;5;246m<int> [39m[23m 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0…
$ PlasticLines    [3m[38;5;246m<int> [39m[23m 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 1, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0…
$ Styrofoam       [3m[38;5;246m<int> [39m[23m 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0…
$ Pellets         [3m[38;5;246m<int> [39m[23m 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0…
$ TotalPlastics   [3m[38;5;246m<int> [39m[23m 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 3, 3, 3, 3, 3, 3, 2, 2, 2, 0, 0, 0, 2, 2, 2, 2, 2, 2, 1, 1, 1, 0, 0, 0, 1, 1, 1, 0, 0…
$ DriftingWood    [3m[38;5;246m<int> [39m[23m 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 1, 1, 3, 3, 3, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 1, 1, 0, 0, 0…
$ Pumice          [3m[38;5;246m<int> [39m[23m 0, 0, 0, 4, 4, 4, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 1, 1, 0, 0, 0, 3, 3, 3, 0, 0…
$ Cs               [3m[38;5;246m<dbl> [39m[23m 691.9619, 691.9619, 691.9619, 566.7916, 566.7916, 566.7916, 653.5089, 653.5089, 653.5089, 950.4462, 950.4462, 950.4462, 2…
$ Date             [3m[38;5;246m<chr> [39m[23m "10/06/2011", "10/06/2011", "10/06/2011", "10/06/2011", "10/06/2011", "10/06/2011", "10/06/2011", "10/06/2011", "10/06/20…
$ VesselTripName  [3m[38;5;246m<chr> [39m[23m "Southern Surveyor/ss2011_t02", "Southern Surveyor/ss2011_t02", "Southern Surveyor/ss2011_t02", "Southern Surveyor/ss2011…
$ CollectedBy     [3m[38;5;246m<chr> [39m[23m "Julia Reisser, Briony Hutton", "Julia Reisser, Briony Hutton", "Julia Reisser, Briony Hutton", "Julia Reisser, Briony Hu…
$ NetType         [3m[38;5;246m<fct> [39m[23m Manta/ 333/ 1 x 0.17, Manta/ 333/ 1 x 0.17, Manta/ 333/ 1 x 0.17, Manta/ 333/ 1 x 0.17, Manta/ 333/ 1 x 0.17, Manta/ 333/…
$ Replicate.y     [3m[38;5;246m<int> [39m[23m 1, 2, 3, 1, 2, 3, 1, 2, 3, 2, 3, 1, 2, 3, 1, 2, 3, 1, 2, 3, 1, 2, 3, 1, 2, 3, 1, 2, 3, 1, 2, 3, 1, 2, 1, 3, 2, 1…
$ StartTime       [3m[38;5;246m<chr> [39m[23m "9:12", "9:32", "9:49", "9:12", "9:32", "9:49", "9:12", "9:32", "9:49", "8:55", "9:11", "5:51", "8:55", "9:11", "5:51", "…
$ EndTime         [3m[38;5;246m<chr> [39m[23m "9:27", "9:47", "10:04", "9:27", "9:47", "10:04", "9:27", "9:47", "10:04", "9:10", "9:26", "6:01", "9:10", "9:26", "6:01"…
$ Duration        [3m[38;5;246m<int> [39m[23m 15, 15, 15, 15, 15, 15, 15, 15, 15, 15, 15, 10, 15, 15, 10, 15, 15, 10, 15, 15, 15, 15, 15, 15, 15, 15, 15, 16, 15, 15, 1…
```

Hide

```
head(finalcombined, n = 10)
```

	NetStation <int>	Replicate.x <int>	HardPlastics <int>	SoftPlastics <int>	PlasticLines <int>	Styrofoam <int>	Pellets <int>
1	1	2	1	0	0	0	0
2	1	2	1	0	0	0	0
3	1	2	1	0	0	0	0
4	1	3	1	0	0	0	0
5	1	3	1	0	0	0	0
6	1	3	1	0	0	0	0
7	1	1	1	0	0	0	0
8	1	1	1	0	0	0	0
9	1	1	1	0	0	0	0
10	2	1	1	0	0	0	0

1-10 of 10 rows | 1-9 of 19 columns

Hide

```
tail(finalcombined, n = 10)
```

	NetStation <int>	Replicate.x <int>	HardPlastics <int>	SoftPlastics <int>	PlasticLines <int>	Styrofoam <int>	Pellets <int>
279	56	3	1	1	0	0	0
280	57	1	7	9	3	0	0
281	57	1	7	9	3	0	0
282	57	1	7	9	3	0	0
283	57	2	1	1	1	0	0
284	57	2	1	1	1	0	0
285	57	2	1	1	1	0	0
286	57	3	6	0	0	0	0
287	57	3	6	0	0	0	0
288	57	3	6	0	0	0	0

1-10 of 10 rows | 1-9 of 19 columns

Tidy & Manipulate Data II

Explanation: In this step, mutate() has been utilised give expand the date variable from one including day, month and year together to 3 seperate variables. This will enable future analyses to be more focused and easily referenced back to the data, especially when investigating factors such as cyclical and seasonal variation. The original date column has been removed to avoid confusion.

Once again, head() has been used to show the updated dataframe.

	Replicate.x <int>	HardPlastics <int>	SoftPlastics <int>	PlasticLines <int>	Styrofoam <int>	Pellets <int>	TotalPlastic <int
1	2	1	0	0	0	0	
2	2	1	0	0	0	0	
3	2	1	0	0	0	0	
4	3	1	0	0	0	0	
5	3	1	0	0	0	0	
6	3	1	0	0	0	0	

6 rows | 1-9 of 21 columns

Scan I

Explanation: This code sums the amount of na values in the final combined dataframe, names it navalues and prints the total number of them (0)

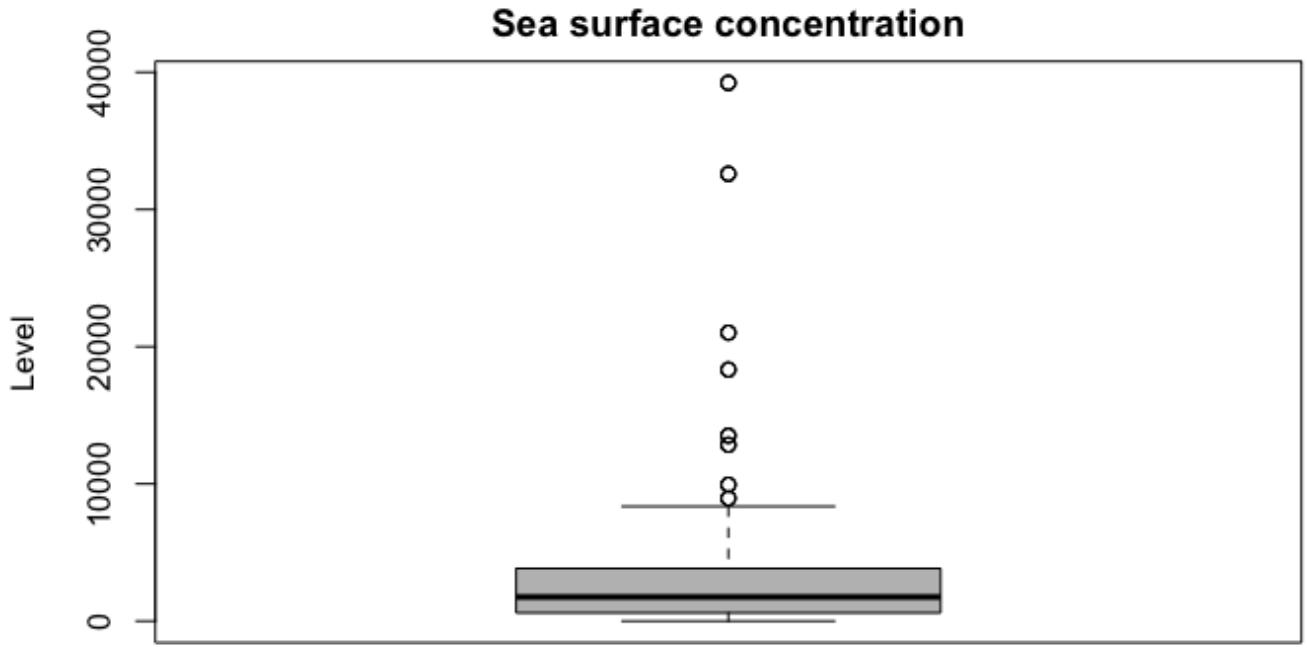
Hide

```
navalues<- sum(is.na(finalcombined))
navalues
```

```
[1] 0
```

Scan II

Explanation: In this step, a box plot has been created from the Cs column of the final combined dataset. The main indicates the title: ‘Sea surface concentration’, the label for the y axis has been donated “level” and the colour of box plot has been denoted grey. The outlier values have then been denoted ‘out’ from the same column and printed to show what they are for the purpose of this report. The corresponding rows for each outlier are then shown to give context of the data.



Hide

```
boxplot(finalcombined$Cs, main="Sea surface concentration",
ylab="Level", col = "grey")
out <- boxplot(finalcombined$Cs, plot=FALSE)$out
print(out)
```

```
 [1] 18324.002 18324.002 18324.002 21021.846 21021.846 21021.846  9943.101  9943.101
9943.101 32595.072 32595.072 32595.072 13518.321 13518.321
[15] 13518.321  8954.088  8954.088  8954.088 39225.930 39225.930 39225.930 12846.118
12846.118 12846.118
```

Hide

```
finalcombined[which(finalcombined$Cs %in% out),]
```

	Replicate.x <int>	HardPlastics <int>	SoftPlastics <int>	PlasticLines <int>	Styrofoam <int>	Pellets <int>	TotalPlas <
91	2	16	0	0	0	0	
92	2	16	0	0	0	0	
93	2	16	0	0	0	0	
109	2	6	0	0	0	0	
110	2	6	0	0	0	0	
111	2	6	0	0	0	0	
112	1	2	0	0	0	0	
113	1	2	0	0	0	0	

	Replicate.x <int>	HardPlastics <int>	SoftPlastics <int>	PlasticLines <int>	Styrofoam <int>	Pellets <int>	TotalPlas <
114	1	2	0	0	0	0	
115	3	11	0	1	0	0	

1-10 of 24 rows | 1-8 of 21 columns Previous **1** 2 3 Next

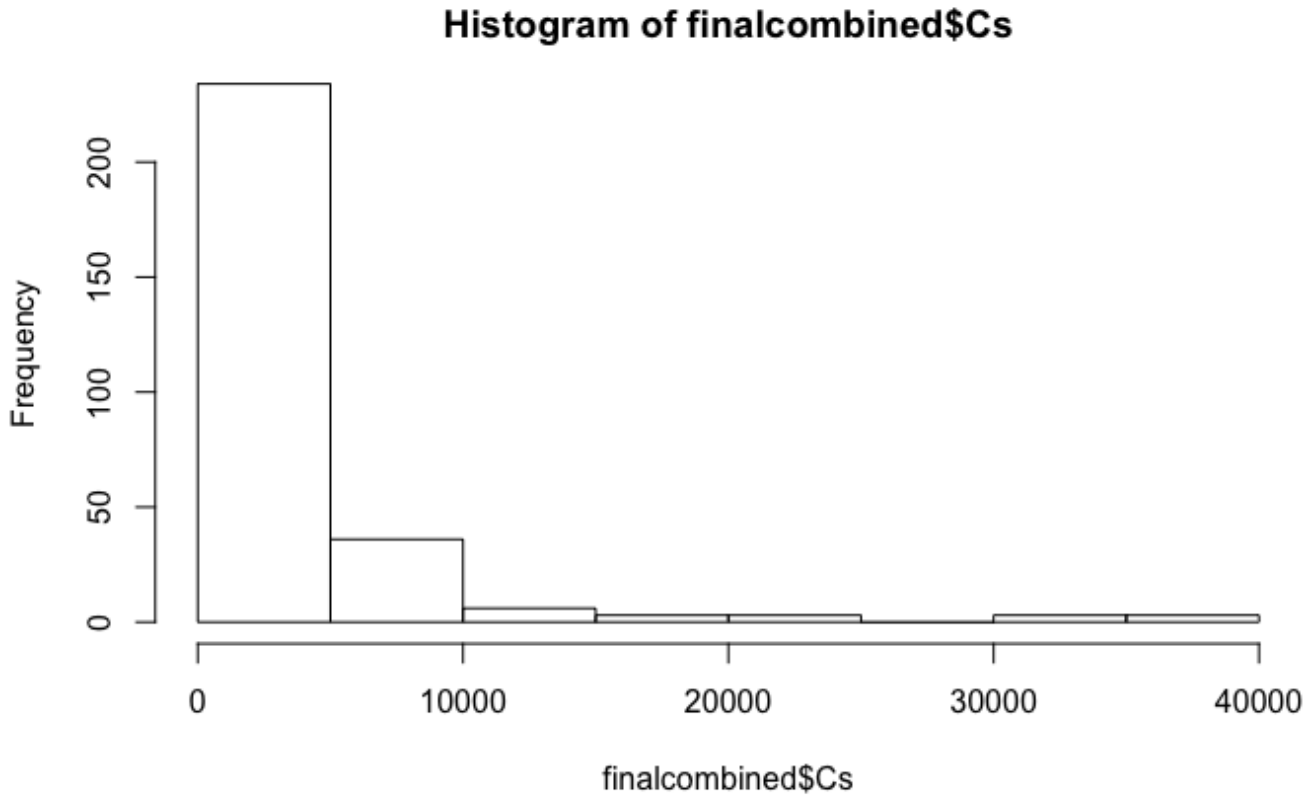
Transform

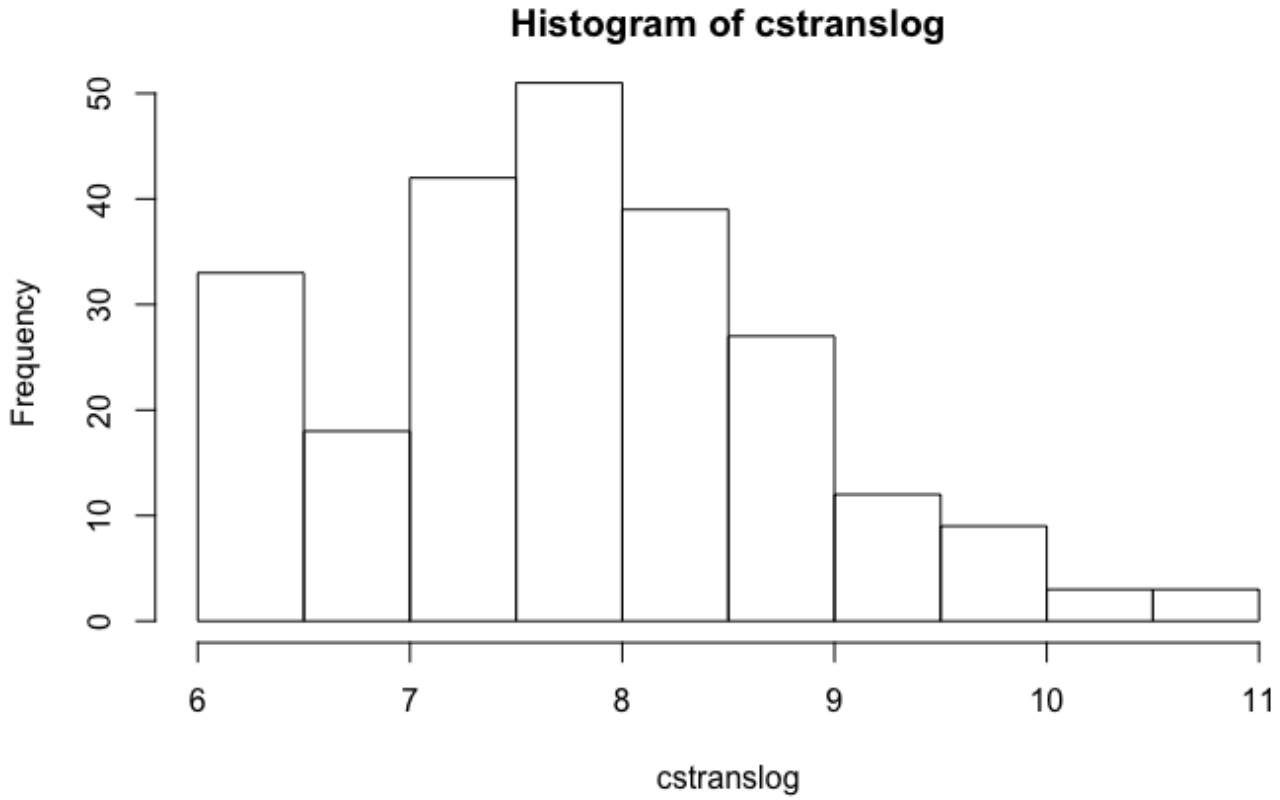
Explanation: A histogram has been created from the Cs column in the final combined graph. No transformations have been shown, these are the original values. The histogram shown is positively skewed. The following code trials different transformations in order to try and normalise the data.

The following transformations have been used:log and log with a base value of 10, square root, the reciprocal and the box cox transformation

After each transformation, a histogram has been created to visualise the data in order to compare the effects.

As shown in the historgrams, the square root, reciprocal transformations showed positively skewed data, whilst box cox showed negatively skewed data. Both log and log base 10 showed more normal distributions.

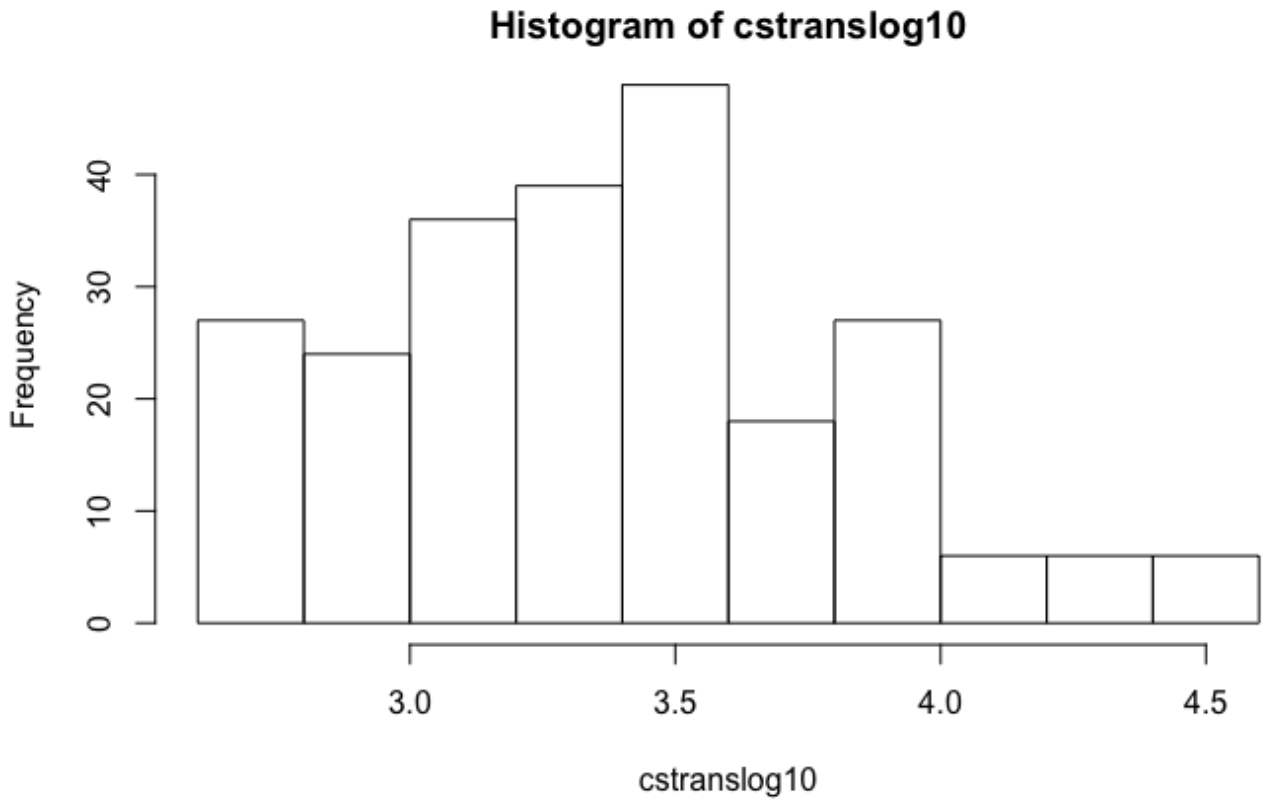




Hide

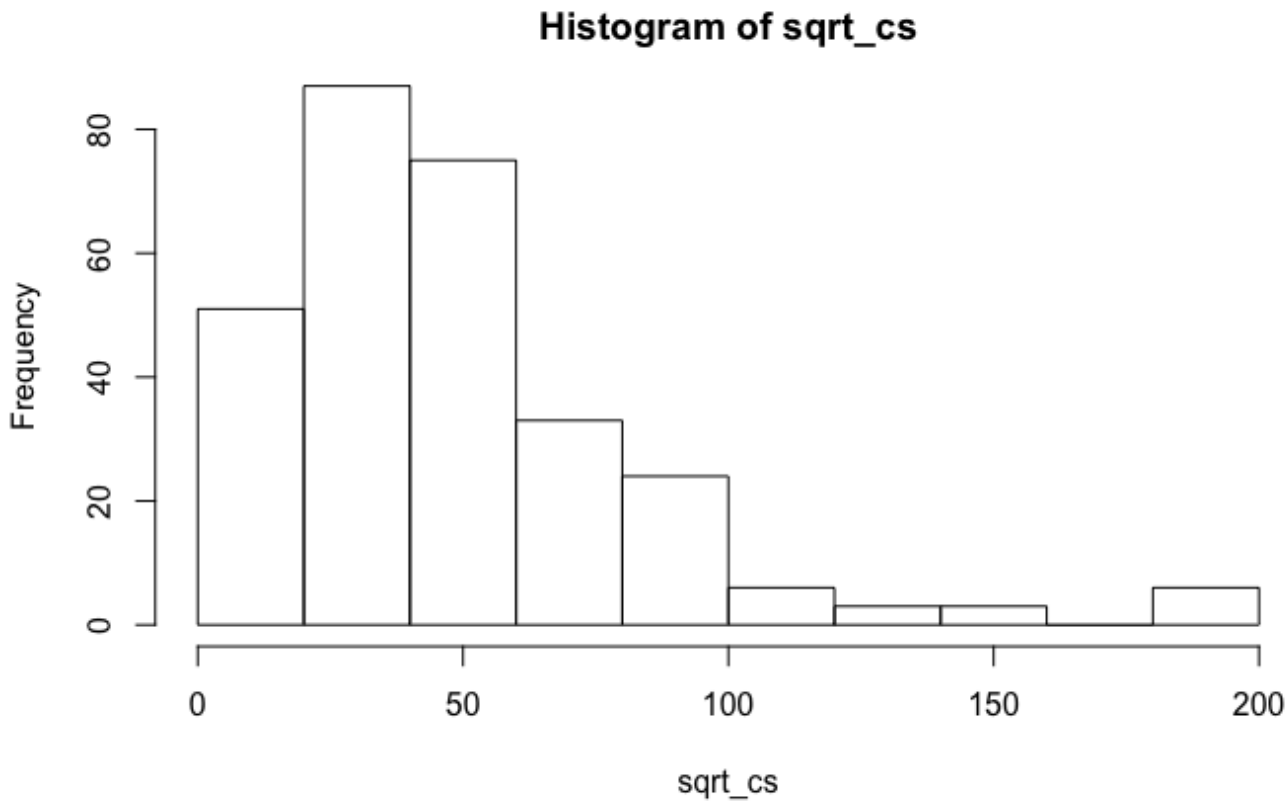
```
hist(finalcombined$Cs)
cstranslog <- finalcombined$Cs
cstranslog <- log(cstranslog)
hist(cstranslog)

cstranslog10 <- finalcombined$Cs
cstranslog10 <- log10(cstranslog10)
hist(cstranslog10)
```



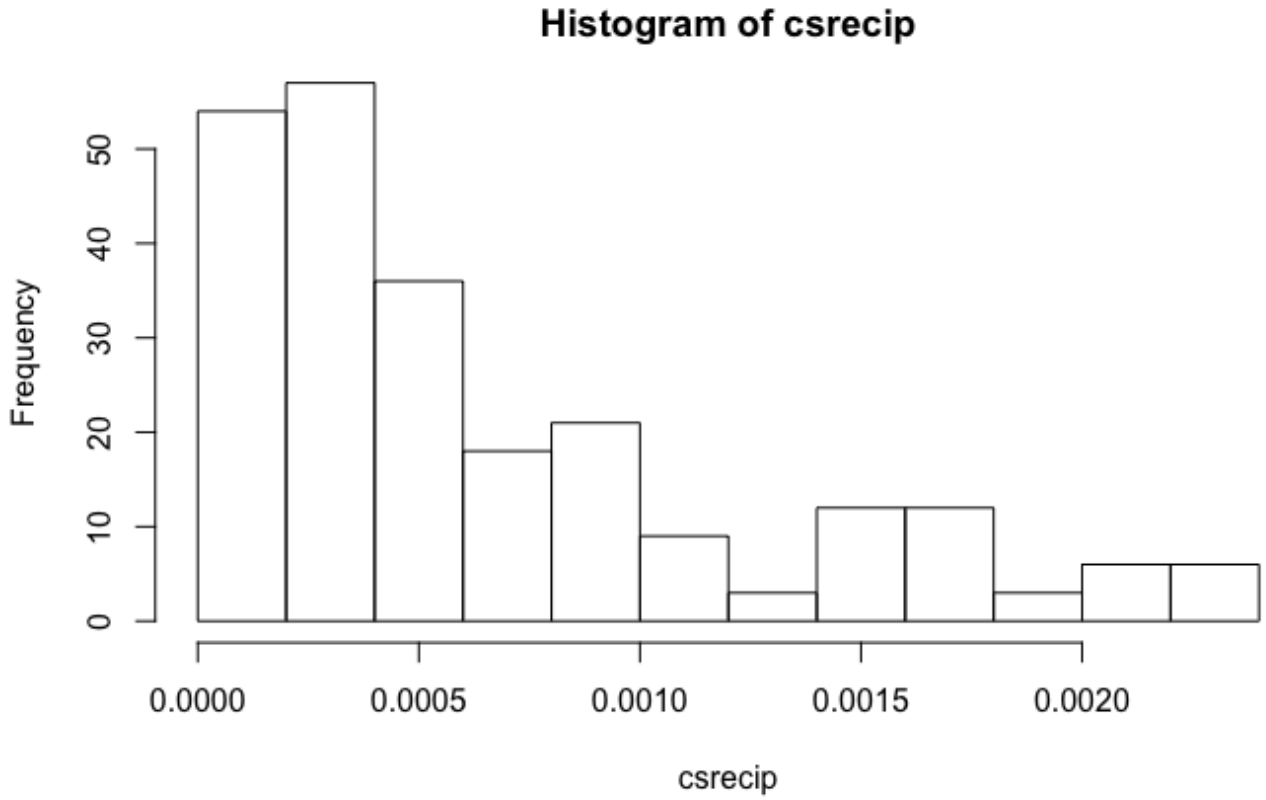
Hide

```
sqrt_cs <- sqrt(finalcombined$Cs)
hist(sqrt_cs)
```



Hide

```
csrecip <- 1/finalcombined$Cs
hist(csrecip)
```



Hide

```
boxcox_cs<- BoxCox(finalcombined$Cs,lambda = "auto")
hist(boxcox_cs)
```