Ocean Data Analysis with R Programming for Early Career Ocean Professionals (ECOPs) (Asia)

Mohamad Lukman Aidid bin Mohd Yusoff

2023-11-05

Individual Project Report

The project will comprise a summary report showing the research question/s, variables analyzed and the results. The report will be 500 words max but the pages will depend on the number of figures generated. Submitted in word or pdf format.

```
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.3     v readr     2.1.4
## v forcats   1.0.0     v stringr   1.5.0
## v ggplot2   3.4.3     v tibble    3.2.1
## v lubridate 1.9.3     v tidyr     1.3.0
## v purrr     1.0.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become error
```

```
library(GGally)
```

```
## Warning: package 'GGally' was built under R version 4.3.2
```

```
## Registered S3 method overwritten by 'GGally':
##   method from
##   +.gg   ggplot2
```

```
library(stats)
library(car)
```

```
## Loading required package: carData
##
## Attaching package: 'car'
##
## The following object is masked from 'package:dplyr':
##
##     recode
```

```
##
## The following object is masked from 'package:purrr':
##
##     some
```

```
library(MASS)
```

```
##
## Attaching package: 'MASS'
##
## The following object is masked from 'package:dplyr':
##
##     select
```

```
library(FactoMineR)
```

```
## Warning: package 'FactoMineR' was built under R version 4.3.2
```

```
library(factoextra)
```

```
## Warning: package 'factoextra' was built under R version 4.3.2

## Welcome! Want to learn more? See two factoextra-related books at https://goo.gl/ve3WBa
```

```
library(fuzzyjoin)
```

```
## Warning: package 'fuzzyjoin' was built under R version 4.3.2
```

For this project, I will be using ship registry data from IRCLASS (Indian Registry of Shipping). In the dataset, you will find data of ships (static info, dimensions, etc.) that have been built and registered under the IRCLASS.

```
irclass <- read.csv(file = "output_irclass.csv")
str(irclass)
```

```
## 'data.frame':    2170 obs. of  64 variables:
##  $ vesselId                   : int  17625 16695 16786 12625 13241 19673 19855 22682 33253 28313 ..
##  $ imoNumber                  : chr  "8022248" "8943648" "8827363" "9082099" ...
##  $ callsign                   : chr  "VWCA" "VTQH" "T2EL5" "AVOY" ...
##  $ officialNumber             : chr  "2786" "2710" "34868917" "3901" ...
##  $ navAids                    : chr  "GC, MC, IMV3, ICS, DSL, TESP, RA, ES, AIS, GPS, NP, NC, MCHB,
##  $ vesselName                 : chr  "RELTUG EIGHT" "M.V. HERMEEZ" "COASTAL LEOPARD" "KANGNA" ...
##  $ homePort                   : chr  "MUMBAI" "MUMBAI" "FUNAFUTI" "MUMBAI" ...
##  $ formerName                 : chr  "ATAMI MARU NO.3" "GE 3,HERMEEZ" "MALAVIYA TWELVE,SKBB KEJAYAA
##  $ flagName                   : chr  "INDIAN" "INDIAN" "Tuvalu" "INDIAN" ...
##  $ dateOfBuild                : chr  "31-Jan-1981" "13-Apr-1998" "31-Jan-1990" "31-Aug-1992" ...
##  $ contractedBuilder          : chr  "YOKOHAMA YACHT CO. LTD., YOKOHAMA, JAPAN" "Bharati Defence &
##  $ dateOfModication           : chr  "" "" "" "" ...
##  $ placeOfbuild               : chr  "YOKOHAMA, JAPAN" "RATNAGIRI,INDIA" "SINGAPORE" "SURAT" ...
##  $ ownerName                  : chr  "RELIANCE INDUSTRIES LTD. , 3rd Floor,Maker Chamber Iv , 222,Na
```

```
##  $ managerName                 : chr  "RELIANCE INDUSTRIES LTD.(SHIP&OFFSHORE DIV) , Shipping & Offsh
##  $ grossTon69                  : num  252 1593 862 1478 225 ...
##  $ netTon69                    : num  76 698 259 551 68 ...
##  $ dwt                         : num  114 2169 885 2277 116 ...
##  $ displacement                : num  412 3131 1420 3074 399 ...
##  $ lwt                         : num  298 962 535 797 283 ...
##  $ lengthOverall               : num  33.3 79 59 77.9 29 ...
##  $ lbp                         : num  29.5 75 56.7 74.6 25.2 ...
##  $ bext                        : num  8.22 14.82 12 NA 0 ...
##  $ bm                          : num  8.2 14.8 12 12.4 8.2 ...
##  $ draught                     : num  2.9 3.29 3 3.5 3.1 ...
##  $ dm                          : num  3.4 4.5 3.8 5 4.07 3.1 4.2 3.05 4.2 19.3 ...
##  $ freeBoard                   : int  518 1232 808 1191 350 1498 1210 850 1100 5319 ...
##  $ hullNotation                : chr  "\"INDIAN COASTAL SERVICE\", TUG, SUL" "\"Indian River Sea Vess
##  $ machineNotation             : chr  "IY" "+ IY" "BWE (s), IY" "IY" ...
##  $ latestSSrecorded            : chr  "SSH03/22" "SSH01/22" "SSH09/14" "SSH08/17" ...
##  $ equipmentLetter             : chr  "24MM" "32MMCC2" "34MMCC2" "32MMCC2" ...
##  $ superStructureDescription   : chr  "F,H,RQDK" "F,H" "BR,RFDK" "F" ...
##  $ riseOffloor                 : chr  "RF 500" "" "" "" ...
##  $ classEntryDate              : chr  "01-Apr-2002" "23-Apr-2003" "05-Oct-2004" "09-Apr-2003" ...
##  $ decksNumber                 : int  1 1 1 1 2 1 1 1 NA 1 ...
##  $ keel                        : chr  "" "" "" "" ...
##  $ bulbousBow                  : chr  "N" "N" "N" "N" ...
##  $ ballastWaterCapacity        : chr  "19.94 m³" "1045.49 m³" "181.79" "1375.91 m³" ...
##  $ framing                     : chr  "TF" "TF IN ER & FPK LF IWO CGO HOLDS" "CF" "TF" ...
##  $ mainClass                   : chr  "-/-" "-/-" "-/BV" "-/ABS" ...
##  $ classStatusString           : chr  "CLASSED" "CLASSED" "SUSPENDED on 07/09/2019" "CLASSED" ...
##  $ bulkheadsFramenumber        : chr  "3 BH AT FR. NOS.13,36,48" "5 BH AT FR. NOS.4,24,68,112,116" "5
##  $ shipType                    : chr  "TUG" "GENERAL CARGO SHIP" "SUPPLY VESSEL" "GENERAL CARGO SHIP
##  $ machineLocation             : chr  "-" "MCHY.AFT" "-" "MCHY.AFT" ...
##  $ passNumber                  : int  0 0 0 0 0 NA NA NA NA NA ...
##  $ holdTanksNumber             : chr  "" "" "" "2HO" ...
##  $ grainBaleLiquidCapacity     : chr  "G.0,B.0,L.0" "G.2593,B.0,L.0" "G.0,B.0,L.0" "G.2100,B.0,L.0"
##  $ insulatedHeatingCoils       : chr  "IN.0" "IN.0" "IN.0" "IN.0" ...
##  $ containerSizeNumber         : int  NA NA NA NA NA NA NA NA NA NA ...
##  $ SWLNumber                   : chr  "2W" "CR 1(10), 0W" "0W" "0W" ...
##  $ hatchways                   : chr  "-" "2HA" "-" "-" ...
##  $ engineDesign                : chr  "NIIGATA" "CUMMINS" "Caterpillar India Pvt. Ltd." "YANMAR" ...
##  $ engineBuild                 : chr  "NIIGATA ENG. CO. LTD." "KIRLOSKAR CUMMINS LTD." "Caterpillar I
##  $ engineModel                 : chr  "6 L 25 CX" "KT 2300M" "CAT 3512" "6N165EN" ...
##  $ dateOfEngineBuild           : chr  "01-Jan-1980" "30-Nov-1996" "01-Jan-1989" "30-Nov-1990" ...
##  $ placeofEngineBuild          : chr  "" "Pune" "Lafayette" "Amagasaki" ...
##  $ engineCycles                : chr  "4SC SA 6CY 250X 320SC SR.GEARED" "4SC SA 12CY 159X 159SC SR.GE
##  $ enginePower                 : chr  "1102KW" "1194KW" "1910KW" "1194KW" ...
##  $ rpm                         : num  NA NA NA NA NA NA 1800 NA 1800 NA ...
##  $ bunkers                     : chr  "36D.0" "81D.0" "149D.0" "74D.0" ...
##  $ boilerHeaterFurnace         : chr  "" "" "" "" ...
##  $ speed                       : num  14.5 10 12.5 11 11 15 8 10.6 8 14 ...
##  $ specialPropellers           : chr  "TF" "TF IN ER & FPK LF IWO CGO HOLDS" "CF" "TF" ...
##  $ auxElectricalGenerationPlant: chr  "GEN 1X 11KW 225V 60HZ AC, GEN 2X 50KW 225V 60HZ AC" "GEN 1X 45
```

For our purposes, I will focus only selected columns comprising vessel dimensions only to set our scope.

```r
colnames(irclass)
```

```
##  [1] "vesselId"                    "imoNumber"
##  [3] "callsign"                    "officialNumber"
##  [5] "navAids"                     "vesselName"
##  [7] "homePort"                    "formerName"
##  [9] "flagName"                    "dateOfBuild"
## [11] "contractedBuilder"           "dateOfModication"
## [13] "placeOfbuild"                "ownerName"
## [15] "managerName"                 "grossTon69"
## [17] "netTon69"                    "dwt"
## [19] "displacement"                "lwt"
## [21] "lengthOverall"               "lbp"
## [23] "bext"                        "bm"
## [25] "draught"                     "dm"
## [27] "freeBoard"                   "hullNotation"
## [29] "machineNotation"             "latestSSrecorded"
## [31] "equipmentLetter"             "superStructureDescription"
## [33] "riseOffloor"                 "classEntryDate"
## [35] "decksNumber"                 "keel"
## [37] "bulbousBow"                  "ballastWaterCapacity"
## [39] "framing"                     "mainClass"
## [41] "classStatusString"           "bulkheadsFramenumber"
## [43] "shipType"                    "machineLocation"
## [45] "passNumber"                  "holdTanksNumber"
## [47] "grainBaleLiquidCapacity"     "insulatedHeatingCoils"
## [49] "containerSizeNumber"         "SWLNumber"
## [51] "hatchways"                   "engineDesign"
## [53] "engineBuild"                 "engineModel"
## [55] "dateOfEngineBuild"           "placeofEngineBuild"
## [57] "engineCycles"                "enginePower"
## [59] "rpm"                         "bunkers"
## [61] "boilerHeaterFurnace"         "speed"
## [63] "specialPropellers"           "auxElectricalGenerationPlant"
```

```r
data <- irclass %>% dplyr::select(vesselId, imoNumber, callsign, vesselName, shipType, lengthOverall, br
str(data)
```

```
## 'data.frame':    1757 obs. of  14 variables:
##  $ vesselId     : int  17625 16695 16786 12625 13241 19673 19855 22682 33253 28313 ...
##  $ imoNumber    : chr  "8022248" "8943648" "8827363" "9082099" ...
##  $ callsign     : chr  "VWCA" "VTQH" "T2EL5" "AVOY" ...
##  $ vesselName   : chr  "RELTUG EIGHT" "M.V. HERMEEZ" "COASTAL LEOPARD" "KANGNA" ...
##  $ shipType     : chr  "TUG" "GENERAL CARGO SHIP" "SUPPLY VESSEL" "GENERAL CARGO SHIP" ...
##  $ lengthOverall: num  33.3 79 59 77.9 29 ...
##  $ bm           : num  8.2 14.8 12 12.4 8.2 ...
##  $ draught      : num  2.9 3.29 3 3.5 3.1 ...
##  $ freeBoard    : int  518 1232 808 1191 350 1498 1210 850 1100 5319 ...
##  $ grossTon69   : num  252 1593 862 1478 225 ...
##  $ netTon69     : num  76 698 259 551 68 ...
##  $ speed        : num  14.5 10 12.5 11 11 15 8 10.6 8 14 ...
##  $ homePort     : chr  "MUMBAI" "MUMBAI" "FUNAFUTI" "MUMBAI" ...
##  $ flagName     : chr  "INDIAN" "INDIAN" "Tuvalu" "INDIAN" ...
```

Exploratory Data Analysis

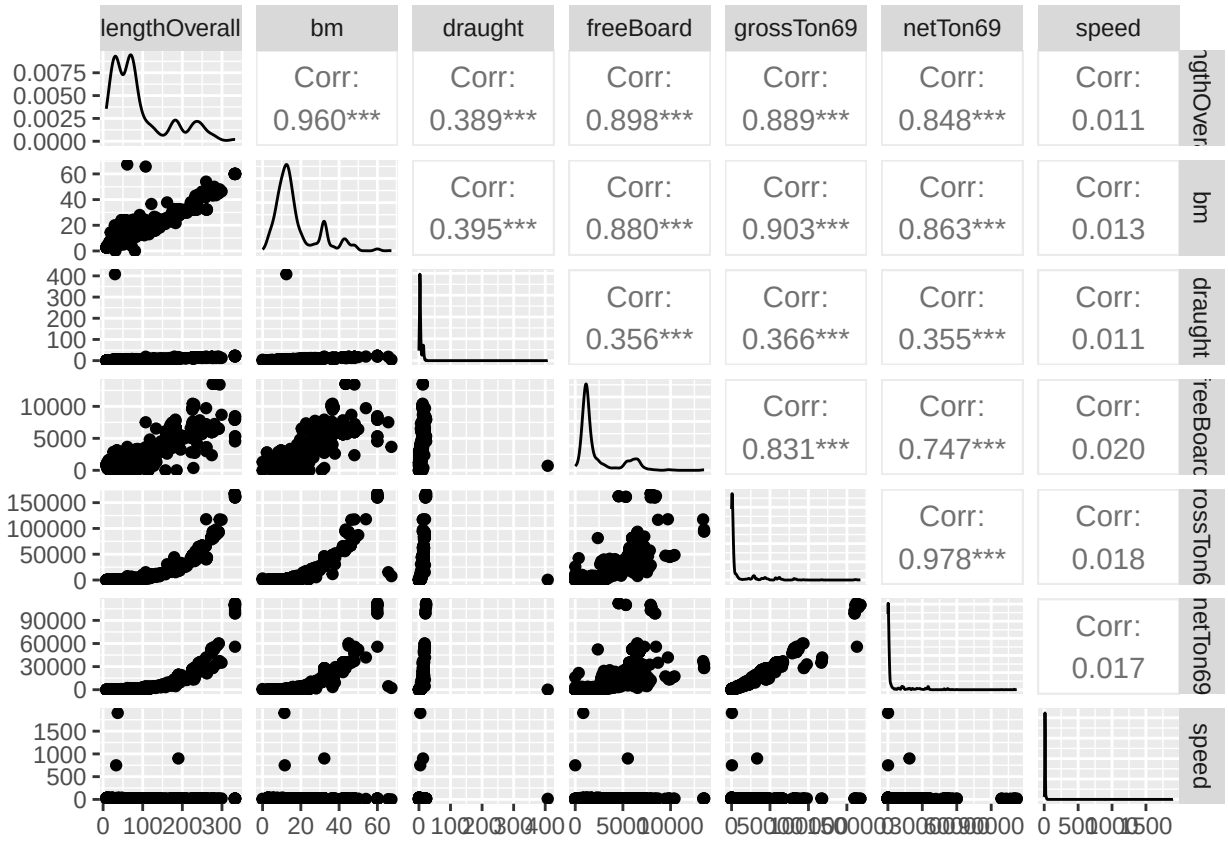
```
head(data)
```

```
##   vesselId imoNumber callsign      vesselName           shipType lengthOverall
## 1    17625   8022248     VWCA    RELTUG EIGHT                TUG         33.30
## 2    16695   8943648     VTQH    M.V. HERMEEZ GENERAL CARGO SHIP         78.95
## 3    16786   8827363    T2EL5 COASTAL LEOPARD      SUPPLY VESSEL         59.00
## 4    12625   9082099     AVOY          KANGNA GENERAL CARGO SHIP         77.90
## 5    13241   8854225     ATPG    ESSAR TUG IV                TUG         29.00
## 6    19673   8026476     VWLE   SUMAI TANGKAS          CREW BOAT         30.15
##     bm draught freeBoard grossTon69 netTon69 speed homePort flagName
## 1  8.2    2.90       518        252       76  14.5   MUMBAI   INDIAN
## 2 14.8    3.29      1232       1593      698  10.0   MUMBAI   INDIAN
## 3 12.0    3.00       808        862      259  12.5 FUNAFUTI   Tuvalu
## 4 12.4    3.50      1191       1478      551  11.0   MUMBAI   INDIAN
## 5  8.2    3.10       350        225       68  11.0   MUMBAI   INDIAN
## 6  6.4    1.60      1498        142       43  15.0   MUMBAI   INDIAN
```

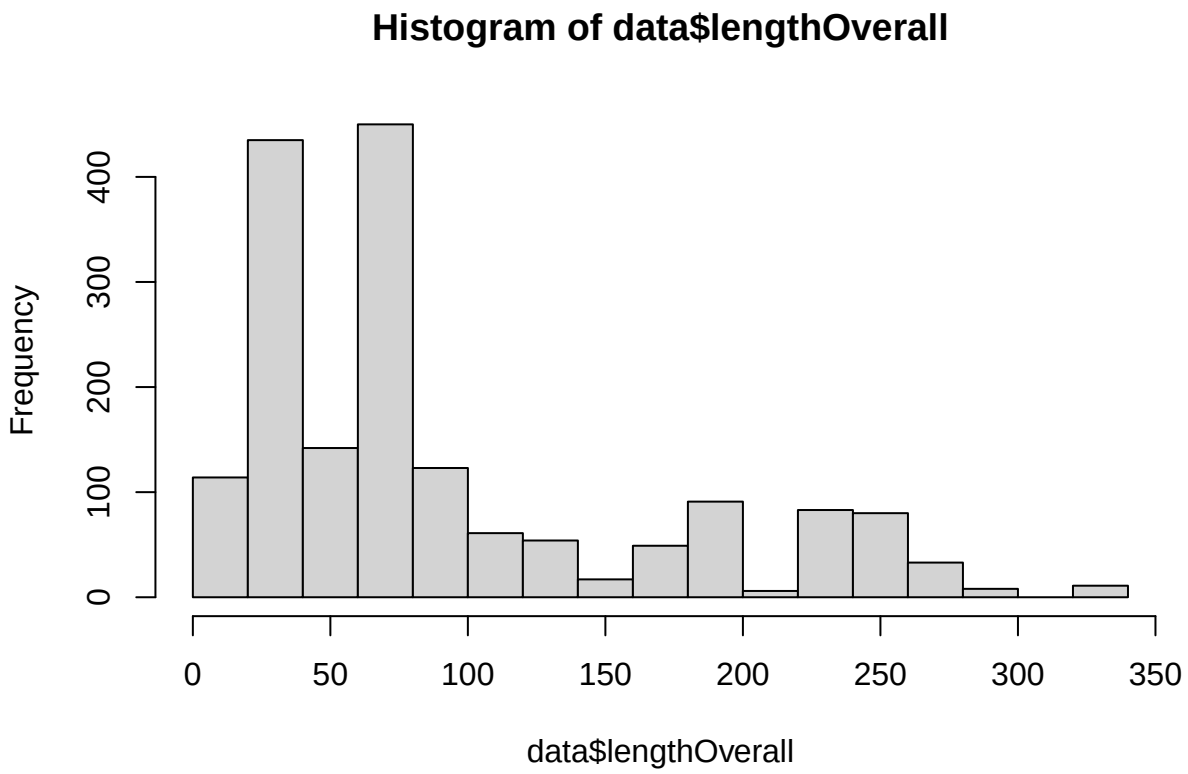
```
summary(data)
```

```
##     vesselId      imoNumber           callsign          vesselName
##  Min.   : 2125   Length:1757        Length:1757        Length:1757
##  1st Qu.:32405   Class :character   Class :character   Class :character
##  Median :46315   Mode  :character   Mode  :character   Mode  :character
##  Mean   :46706
##  3rd Qu.:64717
##  Max.   :69133
##    shipType         lengthOverall          bm           draught
##  Length:1757        Min.   :  8.00   Min.   : 0.00   Min.   :  0.000
##  Class :character   1st Qu.: 33.71   1st Qu.:10.00   1st Qu.:  2.970
##  Mode  :character   Median : 69.80   Median :13.40   Median :  3.900
##                     Mean   : 91.70   Mean   :17.22   Mean   :  5.928
##                     3rd Qu.:117.27   3rd Qu.:20.00   3rd Qu.:  6.549
##                     Max.   :333.15   Max.   :67.33   Max.   :408.000
##    freeBoard       grossTon69          netTon69          speed
##  Min.   :    0   Min.   :     4.46   Min.   :     0   Min.   :   0.00
##  1st Qu.: 1017   1st Qu.:   396.00   1st Qu.:   126   1st Qu.:  10.00
##  Median : 1266   Median :  1342.00   Median :   644   Median :  12.00
##  Mean   : 2292   Mean   : 11242.17   Mean   :  5934   Mean   :  14.68
##  3rd Qu.: 2512   3rd Qu.:  5973.00   3rd Qu.:  2488   3rd Qu.:  14.50
##  Max.   :13516   Max.   :167578.00   Max.   :112240   Max.   :1900.00
##    homePort           flagName
##  Length:1757        Length:1757
##  Class :character   Class :character
##  Mode  :character   Mode  :character
##
##
##
```

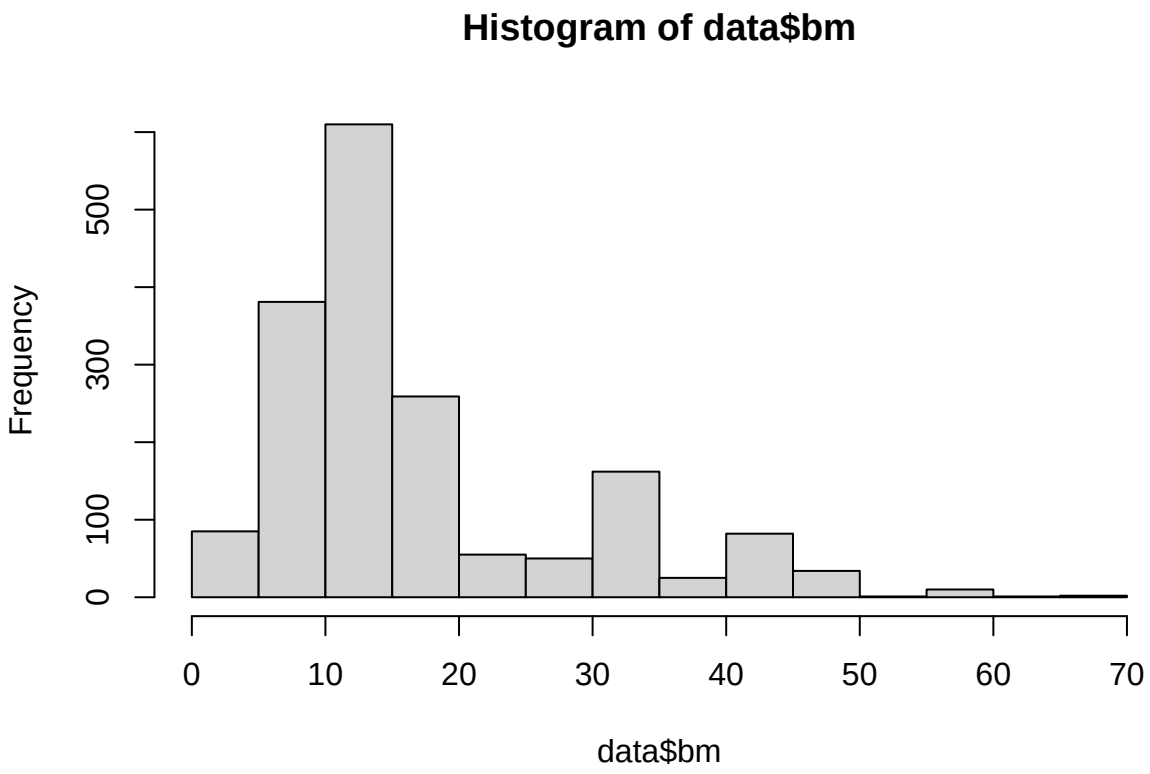
```
ggpairs(select_if(data, is.numeric) %>% dplyr::select(-c(vesselId)))
```



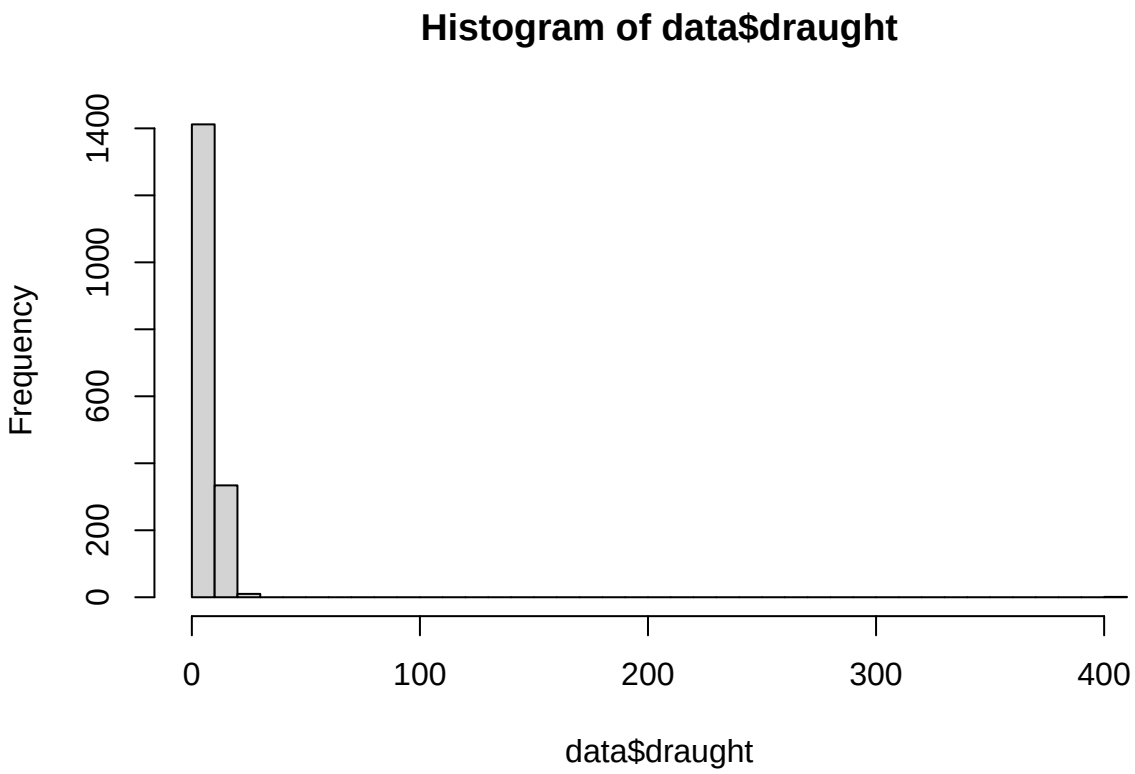

```r
hist(data$lengthOverall)
```


Histogram of data$lengthOverall
Frequency
0
100
200
300
400
0
50
100
150
200
250
300
350
data$lengthOverall

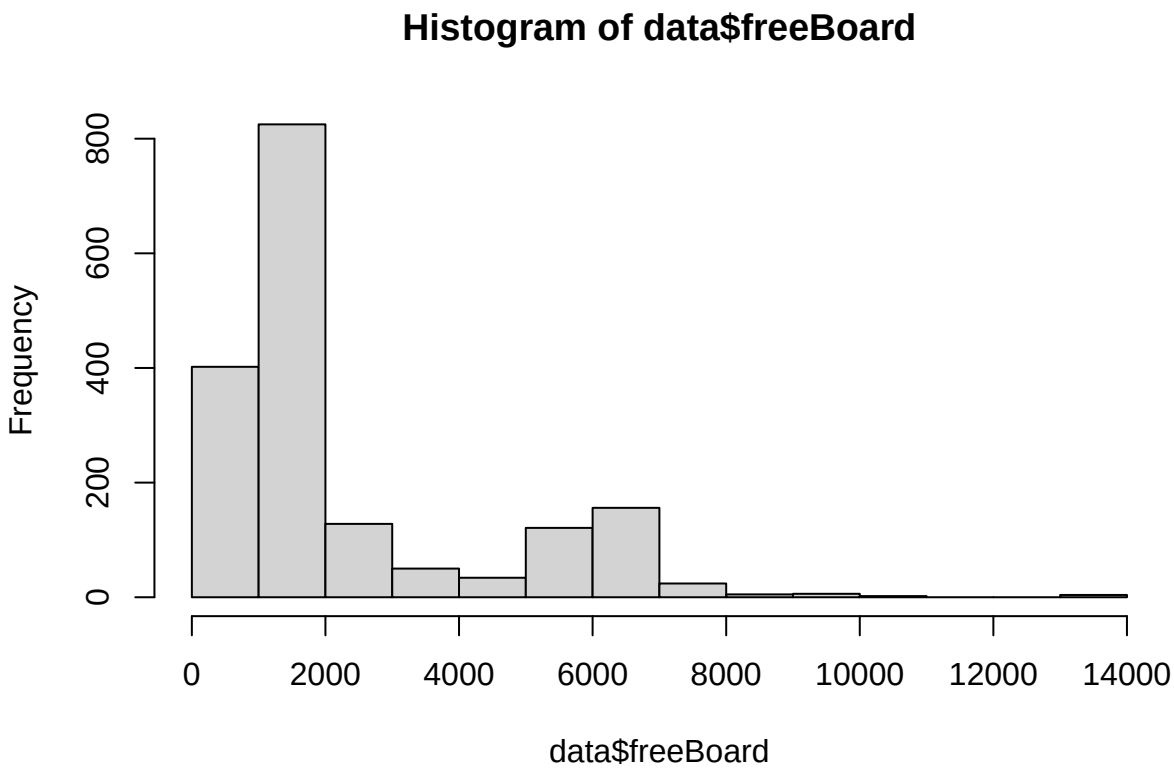
```
hist(data$bm)
```



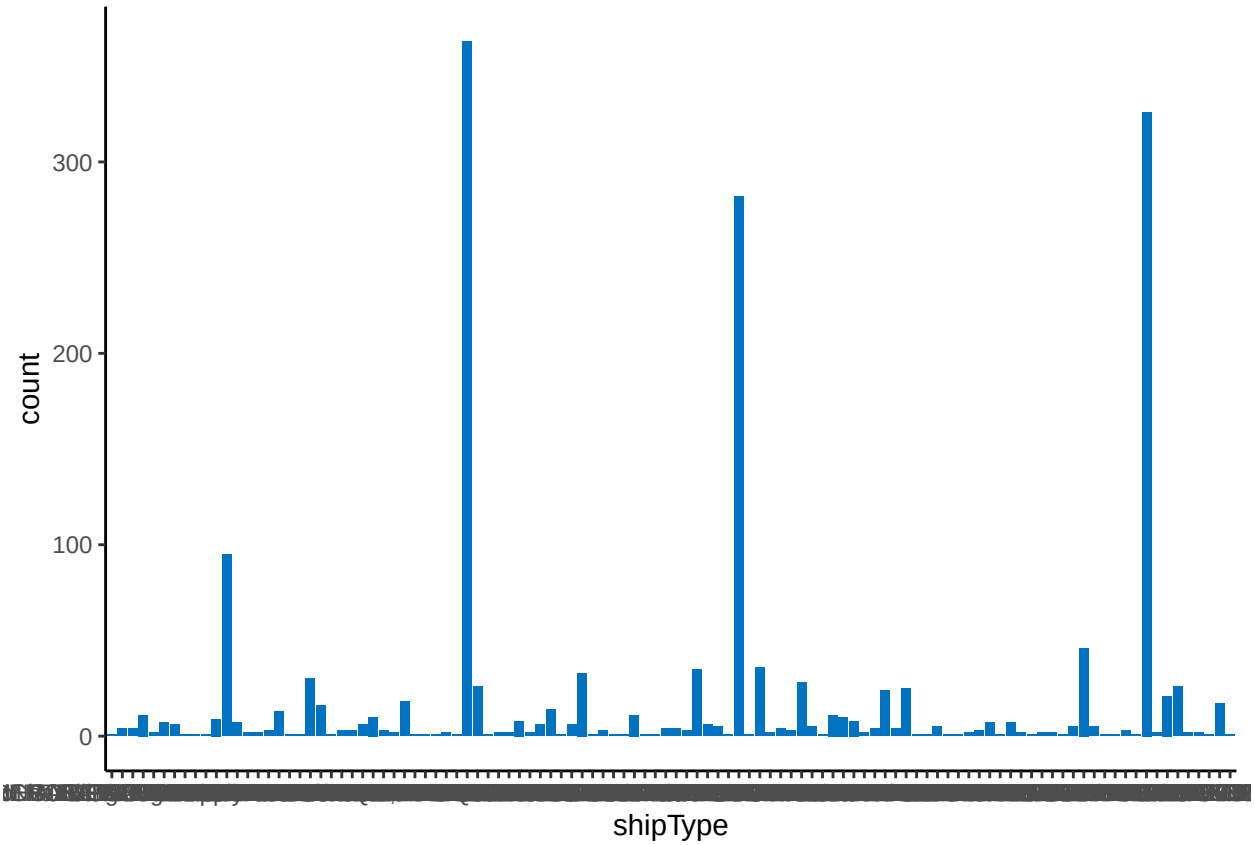

```
hist(data$draught, breaks=50)
```




```
hist(data$freeBoard)
```




```
ggplot(data, aes(shipType)) +
  geom_bar(fill = "#0073C2FF") +
  theme_classic()
```




```
unique(data$shipType)
```

```
##   [1] "TUG"
##   [2] "GENERAL CARGO SHIP"
##   [3] "SUPPLY VESSEL"
##   [4] "CREW BOAT"
##   [5] "OIL TANKER"
##   [6] "BULK CARRIER"
##   [7] "OFFSHORE SUPPORT VESSEL"
##   [8] "ANCHOR HANDLING TUG/ OFFSHORE SUPPORT VESSEL"
##   [9] "CONTAINER SHIP"
##  [10] "TUG/SUPPLY VESSEL"
##  [11] "PASSENGER VESSEL"
##  [12] "UTILITY VESSEL"
##  [13] "LANDING CRAFT"
##  [14] "PASSENGER FERRY"
##  [15] "WORKBOAT"
##  [16] "RO-RO FERRY"
##  [17] "OIL/CHEMICAL TANKER"
##  [18] "LANDING CRAFT ASSAULT"
##  [19] "SURVEY LAUNCH"
##  [20] "RESEARCH VESSEL"
##  [21] "BARGE"
##  [22] "BORDER OUT POST"
##  [23] "CATAMARAN FERRY"
```

```
##  [24] "PATROL VESSEL"
##  [25] "PASSENGER SHIP"
##  [26] "OTHERS"
##  [27] "HOPPER DREDGER"
##  [28] "DRILL SHIP"
##  [29] "PILOT LAUNCH"
##  [30] "RESEARCH GEOLOGICAL SURVEY"
##  [31] "GENERAL DRY CARGO SHIP"
##  [32] "ORE CARRIER"
##  [33] "OIL BARGE"
##  [34] "BUNKER BARGE"
##  [35] "COLUMN STABILISED DRILLING UNIT"
##  [36] "PASSENGER"
##  [37] "FERRY"
##  [38] "ASPHALT CARRIER"
##  [39] "LIQUEFIED GAS CARRIER"
##  [40] "CEMENT CARRIER"
##  [41] "DECK LOADING BARGE"
##  [42] "PATROL CRAFT"
##  [43] "SOLAR ELECTRIC HYBRID PASSENGER FERRY"
##  [44] "LIQUIFIED GAS CARRIER/CHEMICAL TANKER"
##  [45] "HOPPER SUCTION DREDGER"
##  [46] "RESEARCH OCEANOGRAPHIC"
##  [47] "PASSENGER HIGH SPEED CRAFT"
##  [48] "SPLIT HOPPER BARGE"
##  [49] "OFFSHORE SUPPLY VESSEL"
##  [50] "OIL & CHEMICAL TANKER"
##  [51] "LIVESTOCK CARRIER"
##  [52] "TRAILING SUCT HOPPER DREDGE"
##  [53] "DIVING SUPPORT VESSEL"
##  [54] "RO-RO SHIP"
##  [55] "PASSENGER LAUNCH"
##  [56] "PONTOON CRANE"
##  [57] "ANCHOR HANDLING TUG"
##  [58] "SELF ELEVATING DRILLING UNIT"
##  [59] "LAUNCH"
##  [60] "FLOATING CRANE"
##  [61] "PASSENGER/GENERAL CARGO"
##  [62] "DREDGER"
##  [63] "PONTOON"
##  [64] "BARGE DECK LOADING"
##  [65] "MULTIPURPOSE SUPPORT VESSEL"
##  [66] "YACHT"
##  [67] "MOORING LAUNCH"
##  [68] "SELF PROPELLED BARGE"
##  [69] "HOPPER BARGE"
##  [70] "CHEMICAL TANKER"
##  [71] "TUG FIRE FIGHTING"
##  [72] "FAST PATROL VESSEL"
##  [73] "FUEL CUM WATER CARRIER"
##  [74] "PILOT BOAT"
##  [75] "LNG CARRIER"
##  [76] "DECK CARGO SHIP"
##  [77] "MULTIPURPOSE HARBOUR VESSEL"
```

```
##  [78] "WATER TANKER"
##  [79] "RESEARCH CUM FISHING VESSEL"
##  [80] "BARGE CRANE"
##  [81] "UTILITY/CREW BOAT"
##  [82] "Anchor handling Tug Supply Vessel"
##  [83] "LANDING CRAFT MECHANISED"
##  [84] "OFFSHORE PATROL VESSEL"
##  [85] "TRANSHIPMENT CARGO BARGE"
##  [86] "OIL RECOVERY VESSEL"
##  [87] "SELF ELEVATING PLATFORM"
##  [88] "PIPE LAYING/HOOK UP BARGE"
##  [89] "RO-RO CARGO SHIP"
##  [90] "PLATFORM SUPPLY VESSEL"
##  [91] "PATROL BOAT"
##  [92] "TORPEDO LAUNCH AND RECOVERY VESSEL"
##  [93] "ANCHOR HANDLING TUG/ SUPPLY VESSEL"
##  [94] "WELL STIMULATION VESSEL"
##  [95] "BARGE DERRICK/PIPE LAYING"
##  [96] "OIL TANKER FOR CARRIAGE OF ASPHALT"
##  [97] "FLOATING PRODUCTION, STORAGE & OFFLOADING UNIT"
##  [98] "CABLE LAYING VESSEL"
##  [99] "HIGH SPEED PASSENGER VESSEL"
## [100] "HEAVY LIFT/PIPE LAYING SELF PROPELLED VESSEL"
## [101] "MOBILE OFFSHORE DRILLING UNIT"
## [102] "MULTI PURPOSE CARGO CARRIER"
## [103] "CUTTER SUCTION DREDGER"
## [104] "FLOATING PRODUCTION & STORAGE UNIT"
## [105] "BUOY SHIP"
## [106] "ADVERTISEMENT VESSEL"
## [107] "SURVEY VESSEL"
## [108] "GAS CARRIER"
```

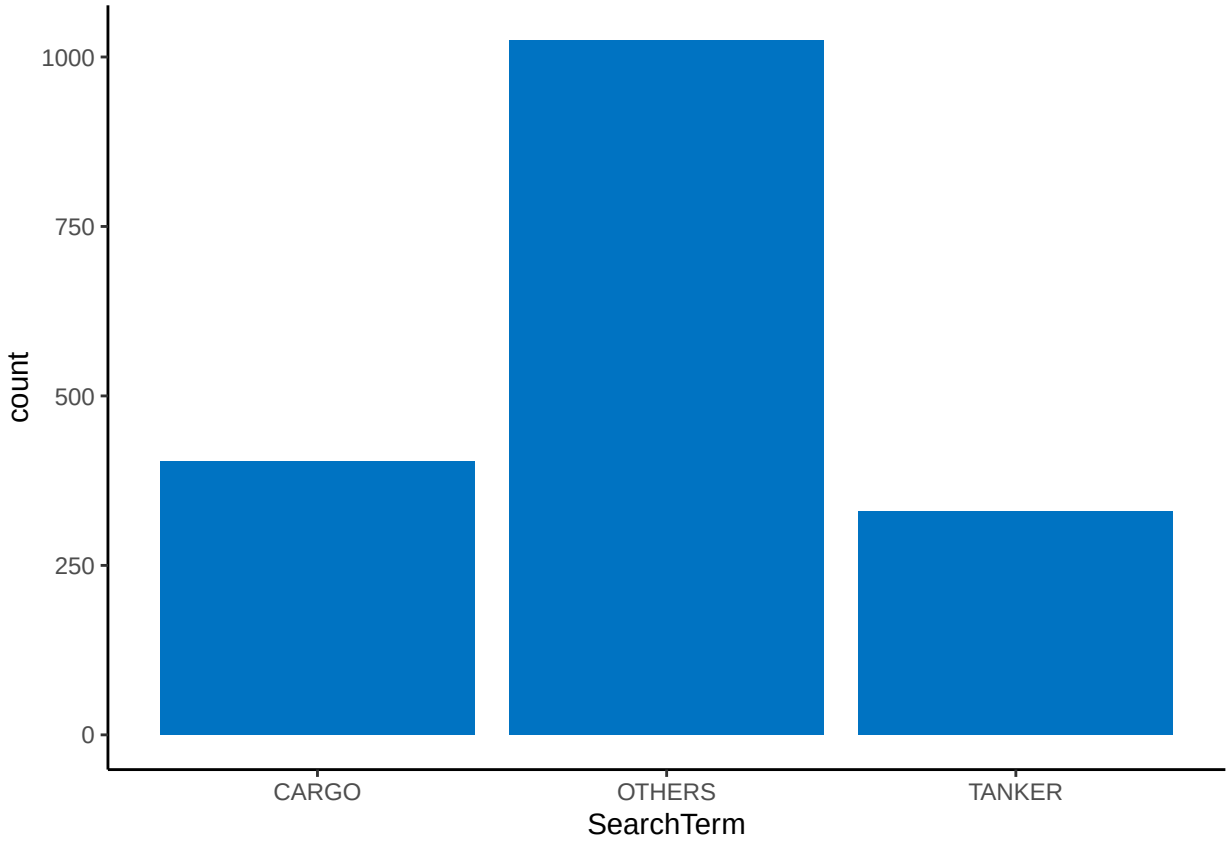
```
CatDF <- data.frame(SearchTerm =  c("CARGO", "TANKER", NA),
                    NewCategory = c("CARGO", "TANKER", "OTHERS"))
data <- regex_left_join(data, CatDF, by = c(shipType  = "SearchTerm"), ignore_case=TRUE) %>% mutate_at(
```

```
head(data)
```

```
##   vesselId imoNumber callsign      vesselName           shipType lengthOverall
## 1    17625   8022248     VWCA    RELTUG EIGHT                TUG         33.30
## 2    16695   8943648     VTQH    M.V. HERMEEZ GENERAL CARGO SHIP         78.95
## 3    16786   8827363    T2EL5 COASTAL LEOPARD      SUPPLY VESSEL         59.00
## 4    12625   9082099     AVOY          KANGNA GENERAL CARGO SHIP         77.90
## 5    13241   8854225     ATPG    ESSAR TUG IV                TUG         29.00
## 6    19673   8026476     VWLE   SUMAI TANGKAS          CREW BOAT         30.15
##     bm draught freeBoard grossTon69 netTon69 speed homePort flagName SearchTerm
## 1  8.2    2.90       518        252       76  14.5   MUMBAI   INDIAN     OTHERS
## 2 14.8    3.29      1232       1593      698  10.0   MUMBAI   INDIAN      CARGO
## 3 12.0    3.00       808        862      259  12.5 FUNAFUTI   Tuvalu     OTHERS
## 4 12.4    3.50      1191       1478      551  11.0   MUMBAI   INDIAN      CARGO
## 5  8.2    3.10       350        225       68  11.0   MUMBAI   INDIAN     OTHERS
## 6  6.4    1.60      1498        142       43  15.0   MUMBAI   INDIAN     OTHERS
##   NewCategory
```

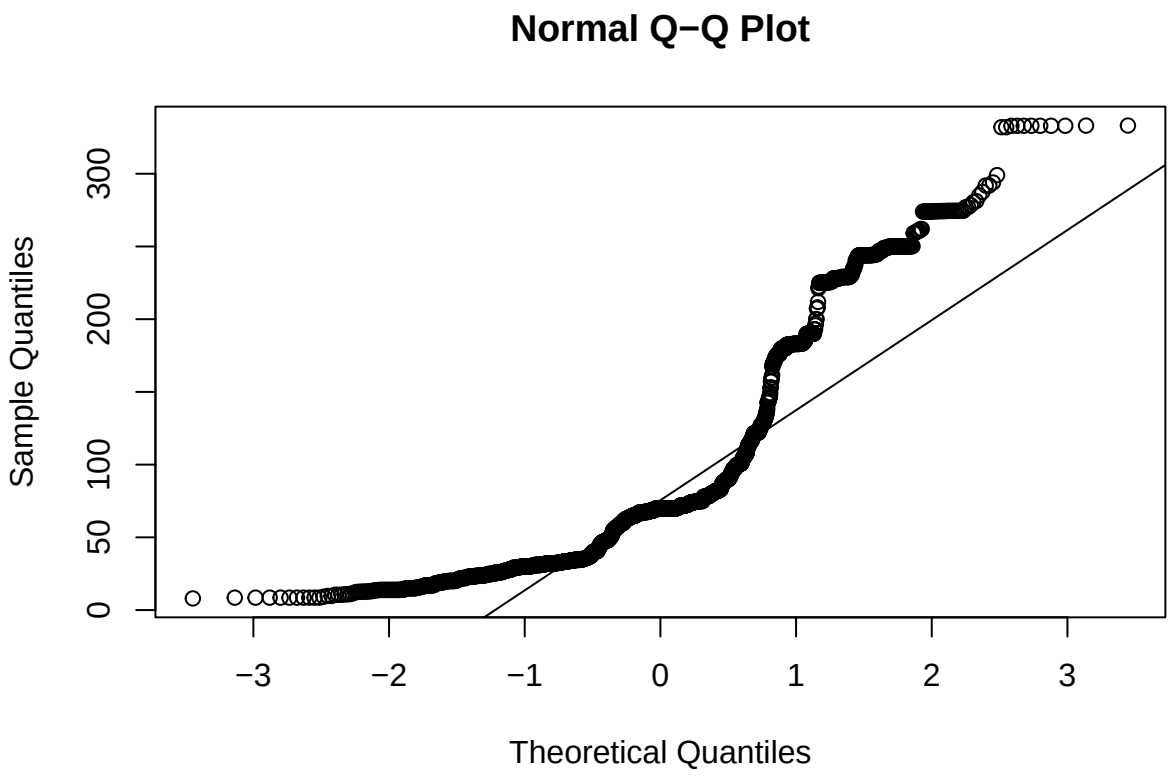
```
## 1      OTHERS
## 2       CARGO
## 3      OTHERS
## 4       CARGO
## 5      OTHERS
## 6      OTHERS
```

```r
ggplot(data, aes(SearchTerm)) +
  geom_bar(fill = "#0073C2FF") +
  theme_classic()
```



Now, we will focus on lengthOverall, as this data generally determines our understanding for the size of a ship.

```
qqnorm(data$lengthOverall)
qqline(data$lengthOverall)
```




```
shapiro.test(data$lengthOverall)
```

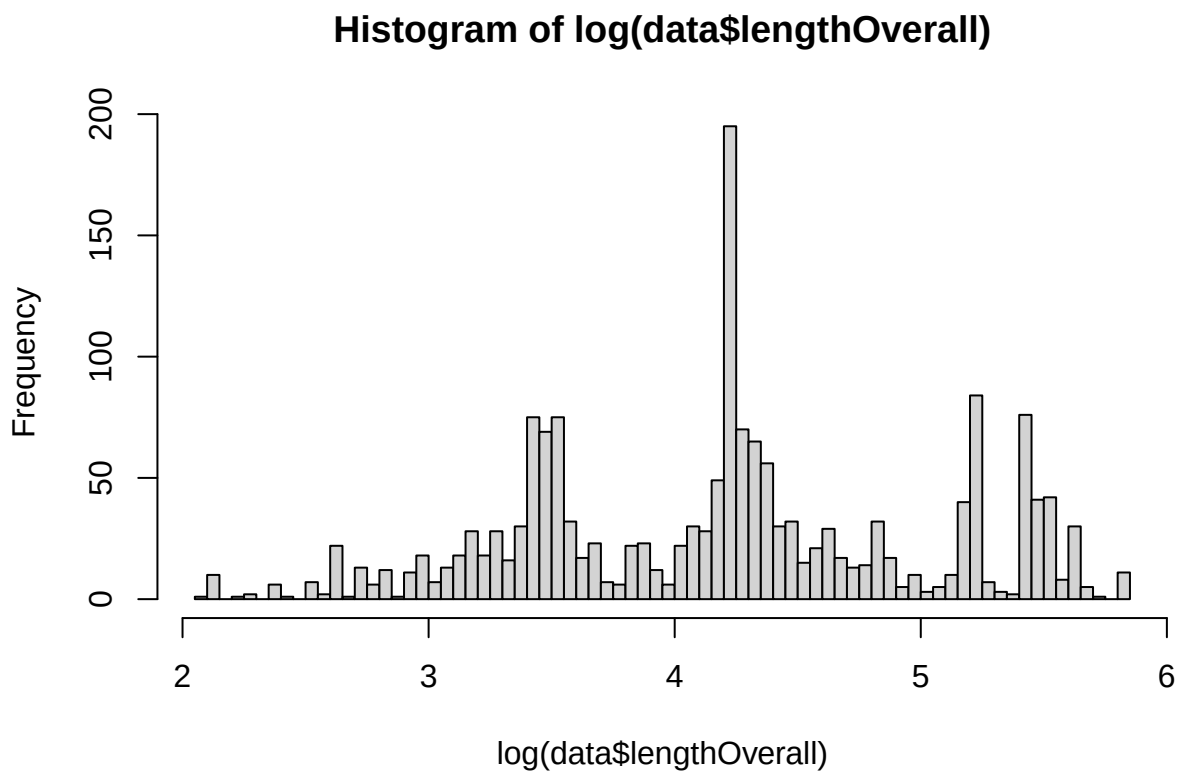
```
##
##  Shapiro-Wilk normality test
##
## data:  data$lengthOverall
## W = 0.83316, p-value < 2.2e-16
```

```
leveneTest(lengthOverall ~ flagName, data)
```

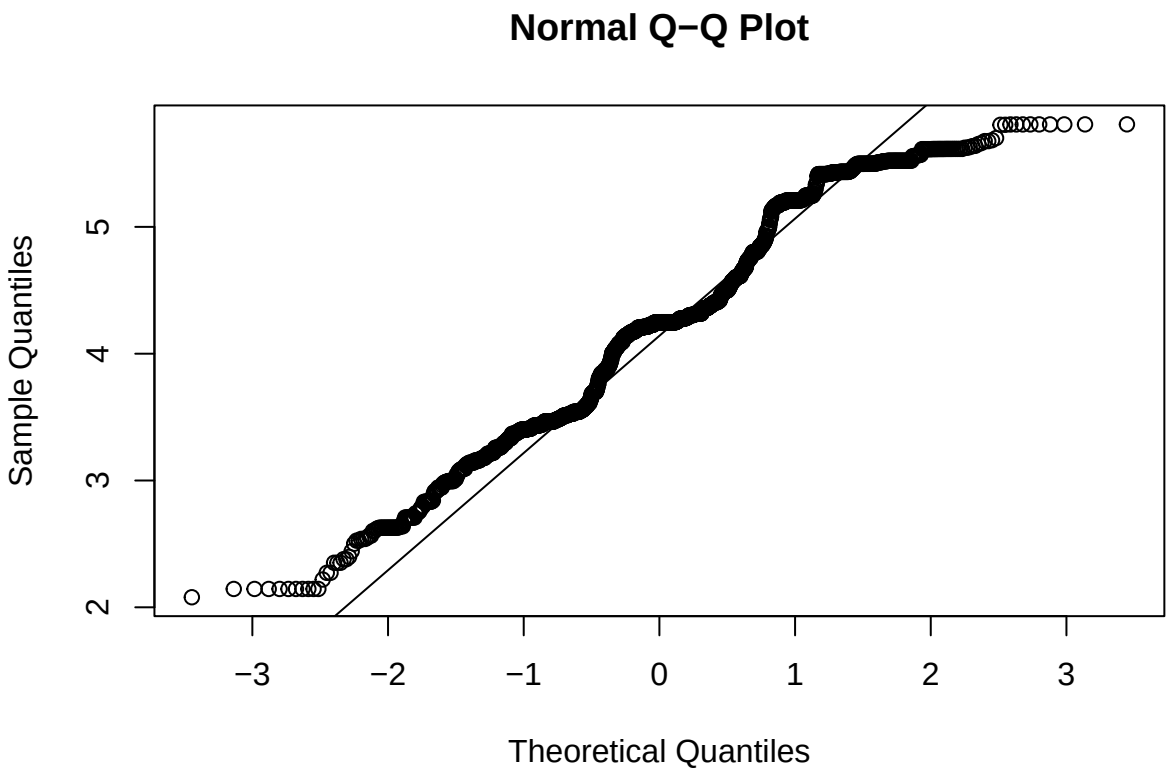
```
## Warning in leveneTest.default(y = y, group = group, ...): group coerced to
## factor.

## Levene's Test for Homogeneity of Variance (center = median)
##         Df F value    Pr(>F)
## group   41  3.0579 4.975e-10 ***
##       1715
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Let's try transforming the data

```
hist(log(data$lengthOverall), 100)
```




```
qqnorm(log(data$lengthOverall))
qqline(log(data$lengthOverall))
```




```
shapiro.test(log(data$lengthOverall))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  log(data$lengthOverall)
## W = 0.97166, p-value < 2.2e-16
```

```
leveneTest(lengthOverall~flagName, data)
```

```
## Warning in leveneTest.default(y = y, group = group, ...): group coerced to
## factor.
```

```
## Levene's Test for Homogeneity of Variance (center = median)
##         Df F value    Pr(>F)
## group   41  3.0579 4.975e-10 ***
##       1715
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Looks like even though the transformed data look much better, it is still not enough to satisfy normality and homoscedasticity.

```r
unique(data$flagName)
```

```
##  [1] "INDIAN"                          "Tuvalu"
##  [3] "Sri Lanka"                       "U.A.E."
##  [5] "Republic Of Vanuatu"             "Panama"
##  [7] "Mauritius"                       "Palau"
##  [9] "Bangladesh"                      ""
## [11] "St. Kitts and Nevis"             "Liberia"
## [13] "Singapore"                       "-"
## [15] "Philippines"                     "Cook Islands"
## [17] "Maldives"                        "Indonesia"
## [19] "Gabon"                           "Cameroon"
## [21] "Barbados"                        "Myanmar"
## [23] "St. Vincent and the Grenadines" "Malta"
## [25] "Bulgaria"                        "Antigua and Barbuda"
## [27] "Belize"                          "Vietnam"
## [29] "Niue"                            "Republic Of Equatorial Guinea"
## [31] "Comoros"                         "Greece"
## [33] "Cyprus"                          "Sultanate Of Oman"
## [35] "Marshall Islands"                "Bahamas"
## [37] "Uganda"                          "Thailand"
## [39] "Commonwealth of Dominica"        "Guinea-Bissau"
## [41] "Qatar"                           "Turkey"
```

The data for 'lengthOverall' is not normal and could not be transformed into a normal distribution via log transform, as both data did not pass the Shapiro-Wilk test .

Comparing variances of 'lengthOverall' and 'flagName' via the Levene test also indicated that the data set is not homoscedastic.

Therefore, ANOVA would not be suitable to test this data. The most appropriate test for this type of data would be the Kruskal-Wallis test.

```r
kruskal.test(lengthOverall~flagName, data=data)
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  lengthOverall by flagName
## Kruskal-Wallis chi-squared = 462.43, df = 41, p-value < 2.2e-16
```

Modeling - we will try to model lengthOverall from the other variables

```
str(data)
```

```
## 'data.frame':    1757 obs. of  16 variables:
##  $ vesselId     : int  17625 16695 16786 12625 13241 19673 19855 22682 33253 28313 ...
##  $ imoNumber    : chr  "8022248" "8943648" "8827363" "9082099" ...
##  $ callsign     : chr  "VWCA" "VTQH" "T2EL5" "AVOY" ...
##  $ vesselName   : chr  "RELTUG EIGHT" "M.V. HERMEEZ" "COASTAL LEOPARD" "KANGNA" ...
##  $ shipType     : chr  "TUG" "GENERAL CARGO SHIP" "SUPPLY VESSEL" "GENERAL CARGO SHIP" ...
##  $ lengthOverall: num  33.3 79 59 77.9 29 ...
##  $ bm           : num  8.2 14.8 12 12.4 8.2 ...
##  $ draught      : num  2.9 3.29 3 3.5 3.1 ...
##  $ freeBoard    : int  518 1232 808 1191 350 1498 1210 850 1100 5319 ...
##  $ grossTon69   : num  252 1593 862 1478 225 ...
##  $ netTon69     : num  76 698 259 551 68 ...
##  $ speed        : num  14.5 10 12.5 11 11 15 8 10.6 8 14 ...
##  $ homePort     : chr  "MUMBAI" "MUMBAI" "FUNAFUTI" "MUMBAI" ...
##  $ flagName     : chr  "INDIAN" "INDIAN" "Tuvalu" "INDIAN" ...
##  $ SearchTerm   : chr  "OTHERS" "CARGO" "OTHERS" "CARGO" ...
##  $ NewCategory  : chr  "OTHERS" "CARGO" "OTHERS" "CARGO" ...
```

```
bm_only = glm(lengthOverall~bm, data = data, "gaussian")
draught_only = glm(lengthOverall~draught, data = data, "gaussian")
freeboard_only = glm(lengthOverall~freeBoard, data = data, "gaussian")
grossTon69_only = glm(lengthOverall~grossTon69, data = data, "gaussian")
netTon69_only = glm(lengthOverall~netTon69, data = data, "gaussian")
all_model = glm(lengthOverall~bm+draught+freeBoard+grossTon69+netTon69, data = data, "gaussian")
summary(all_model)
```

```
##
## Call:
## glm(formula = lengthOverall ~ bm + draught + freeBoard + grossTon69 +
##     netTon69, family = "gaussian", data = data)
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -8.1966114  1.2259791  -6.686 3.08e-11 ***
## bm           4.4415246  0.1095866  40.530  < 2e-16 ***
## draught      0.0464616  0.0455983   1.019    0.308
## freeBoard    0.0096704  0.0005357  18.053  < 2e-16 ***
## grossTon69  -0.0006112  0.0001357  -4.503 7.14e-06 ***
## netTon69     0.0013203  0.0001966   6.715 2.53e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian family taken to be 345.7498)
##
##     Null deviance: 9653105  on 1756  degrees of freedom
## Residual deviance:  605408  on 1751  degrees of freedom
## AIC: 15265
##
## Number of Fisher Scoring iterations: 2
```

```
draught_only_2 = glm(lengthOverall~draught, data = data, Gamma(link = log))
```

```
## Warning: glm.fit: algorithm did not converge
```

```
grossTon69_only_2 = glm(lengthOverall~grossTon69, data = data, Gamma(link = log))
all_model_2 = glm(lengthOverall~bm+draught+freeBoard+grossTon69+netTon69, data = data, Gamma(link = log
summary(all_model_2)
```

```
##
## Call:
## glm(formula = lengthOverall ~ bm + draught + freeBoard + grossTon69 +
##     netTon69, family = Gamma(link = log), data = data)
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.755e+00  2.108e-02 130.673   <2e-16 ***
## bm           9.803e-02  1.885e-03  52.016   <2e-16 ***
## draught     -5.149e-04  7.842e-04  -0.657   0.5115
## freeBoard    2.061e-05  9.212e-06   2.238   0.0254 *
## grossTon69  -2.257e-05  2.334e-06  -9.669   <2e-16 ***
## netTon69     3.407e-06  3.381e-06   1.008   0.3138
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for Gamma family taken to be 0.1022622)
##
##     Null deviance: 1107.77  on 1756  degrees of freedom
## Residual deviance:  179.16  on 1751  degrees of freedom
## AIC: 15789
##
## Number of Fisher Scoring iterations: 7
```

Looks like the best predictor is draught.

We can compare models with AIC.

```
AIC(bm_only,draught_only,draught_only_2,freeboard_only,grossTon69_only,grossTon69_only_2,netTon69_only,a
```

```
##                   df      AIC
## bm_only            3 15632.17
## draught_only       3 19834.57
## draught_only_2     3 17214.43
## freeboard_only     3 17244.29
## grossTon69_only    3 17374.74
## grossTon69_only_2  3 17599.83
## netTon69_only      3 17889.41
## all_model          7 15265.06
## all_model_2        7 15789.09
```

From AIC, looks like the all_model is best, although all of them have very high AIC score. So the modeling is probably dubious for this data set. Let's check the residuals...

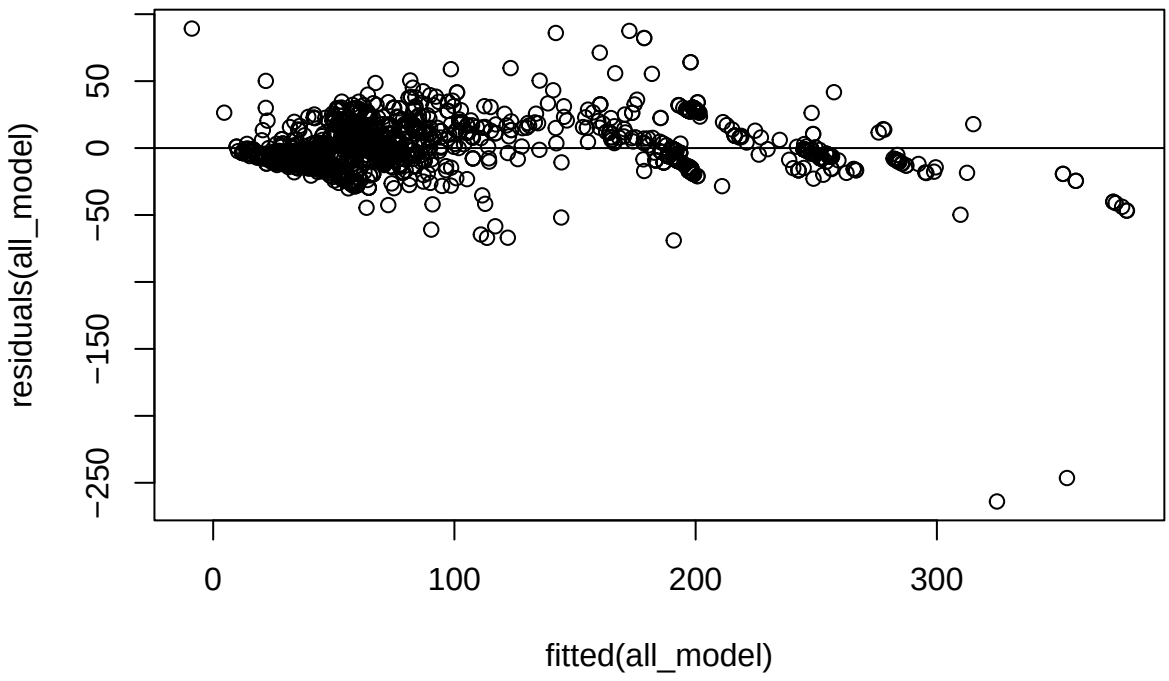
```
shapiro.test(residuals(all_model))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  residuals(all_model)
## W = 0.8329, p-value < 2.2e-16
```

```
shapiro.test(residuals(bm_only))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  residuals(bm_only)
## W = 0.78395, p-value < 2.2e-16
```

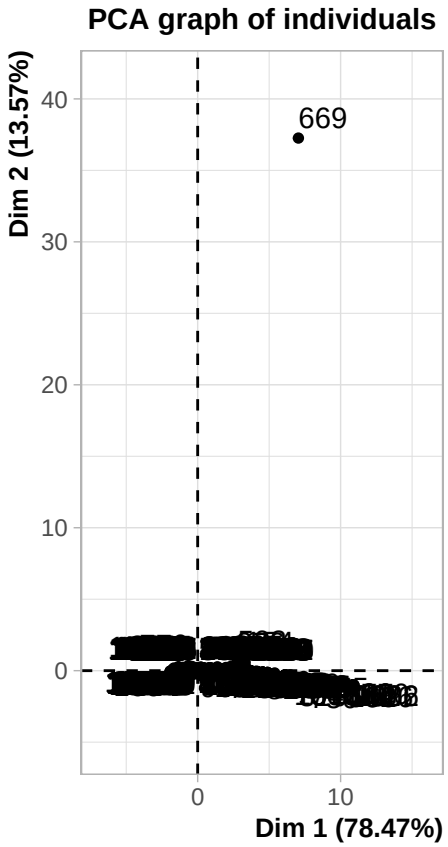
```
plot(residuals(all_model)~fitted(all_model))
abline(h=0)
```

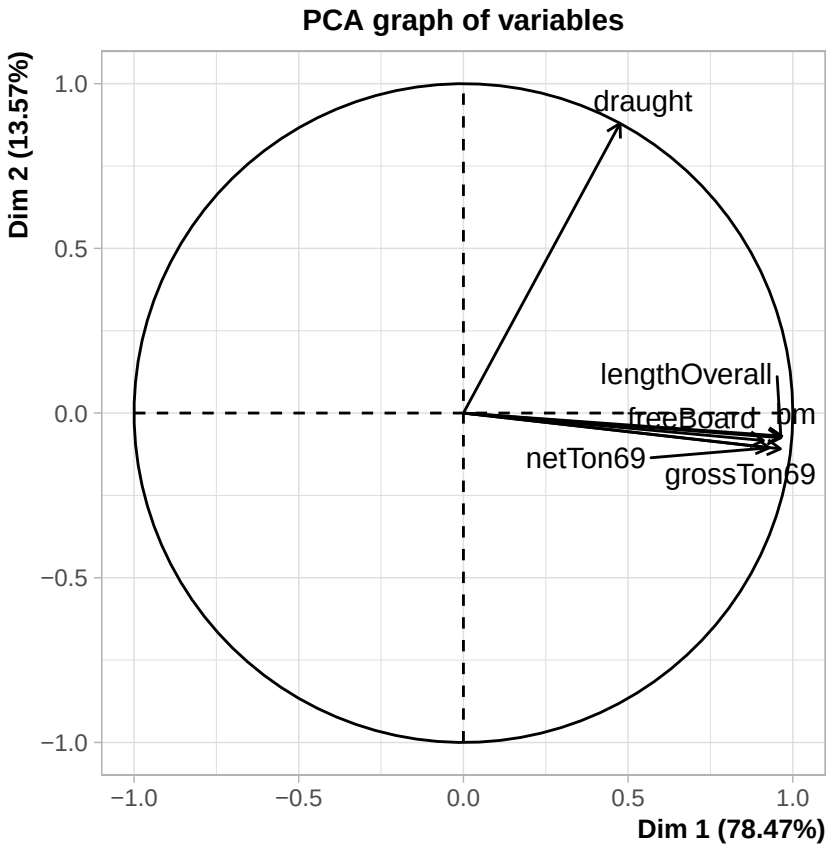


Though the residuals are not normally distributed, they look quite random in the fitted plot.

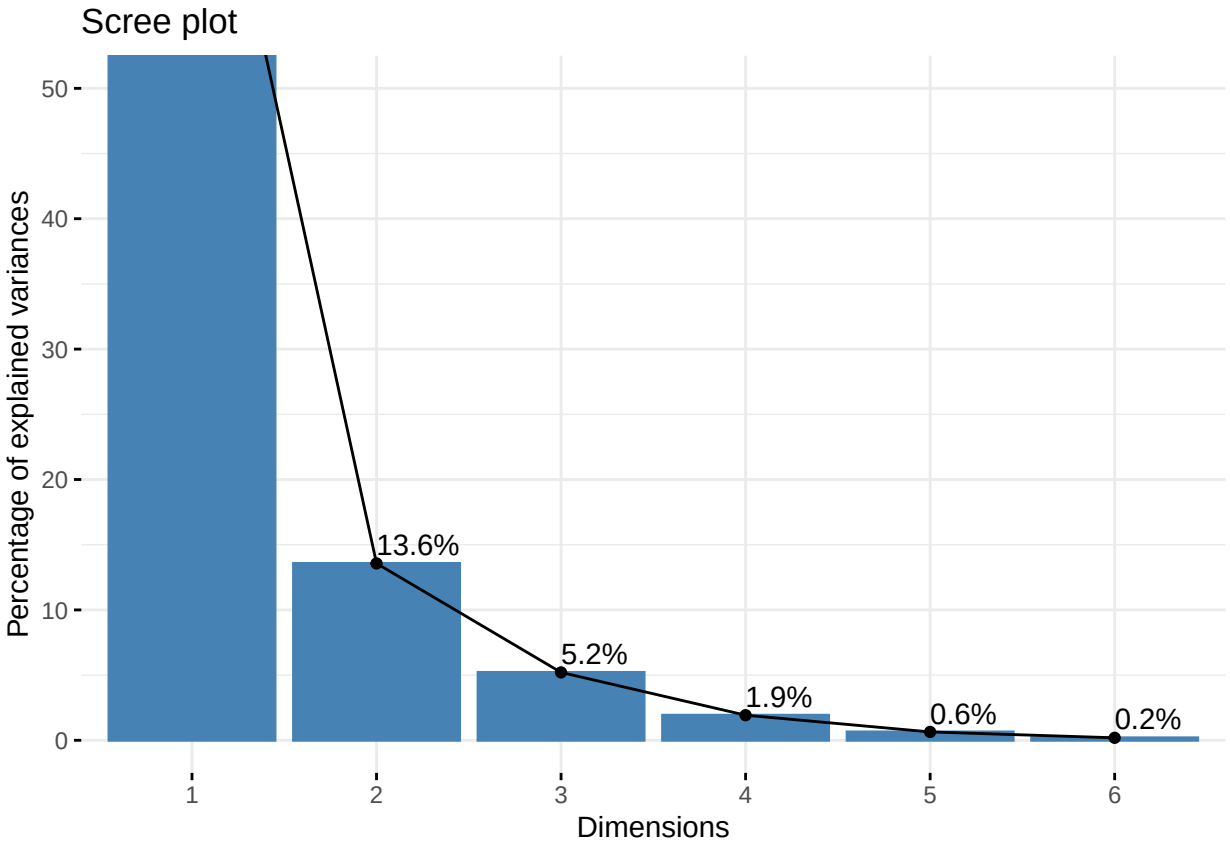
Finally, we will do PCA.

```
pca_results <- PCA(data %>%
                     dplyr::select(lengthOverall,
                     bm,
                     draught,
                     netTon69,
                     grossTon69,
                     freeBoard))
```



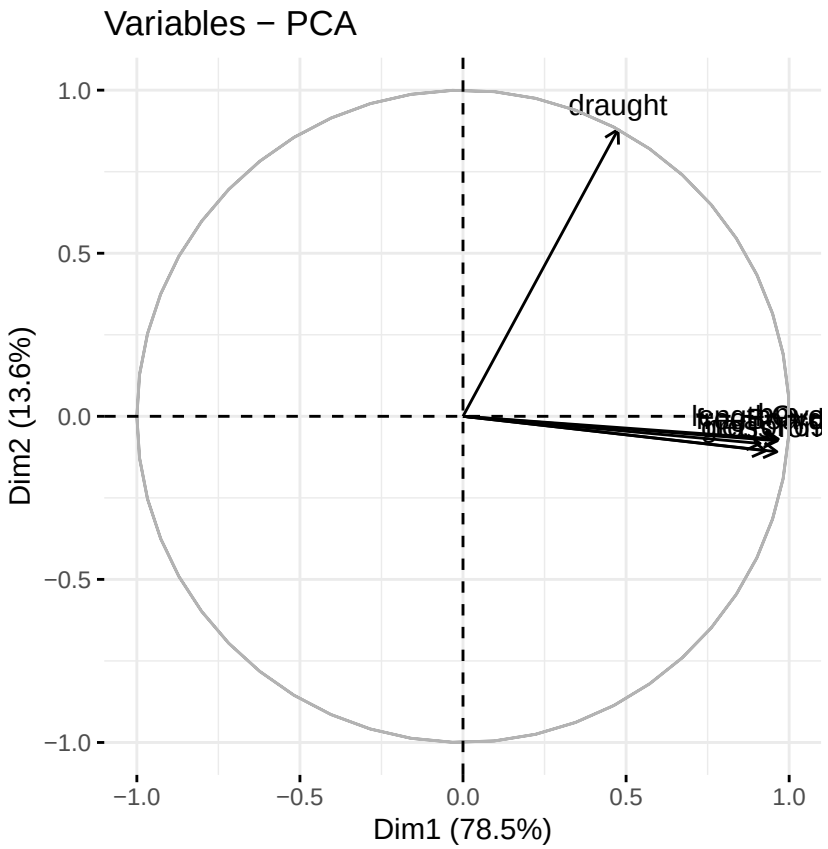



```
fviz_eig(pca_results, addlabels = TRUE, ylim = c(0, 50))
```



Percentage of total variance explained by the first two principal components is ~64%.

```
fviz_pca_var(pca_results)
```



From the above biplot, all of the variables considered except for draught are very tightly correlated to each other and positively correlated to Dim1. These variables are not at all correlated to Dim2.

Only draught is quite positively correlated with Dim2 and it is also slightly positively correlated with Dim1.

Conclusion:

Standard statistical tests are inconclusive in results pertaining this dataset. However, PCA shows that some of the variables are very tightly correlated to one another. These are the traditional dimensions of a ship, namely lengthOverall (length), beam (bm; width), freeBoard (proxy for height), and weight (both gross and net tonnage).