ADAPT Program Day 2

# Loading packages

library(readxl, quietly=T)  
library(dplyr, quietly=T)

##   
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':  
##   
## filter, lag

## The following objects are masked from 'package:base':  
##   
## intersect, setdiff, setequal, union

library(ggplot2, quietly=T)  
library(reshape2, quietly=T)  
library(stringr, quietly=T)  
library(lubridate, quietly=T)

##   
## Attaching package: 'lubridate'

## The following object is masked from 'package:base':  
##   
## date

# Reading files

## Getting the number of files

filenames <- dir(path="./data", pattern="^ADAPT\_Sample\_dataset.\\.csv$")  
  
print(filenames)

## [1] "ADAPT\_Sample\_dataset1.csv" "ADAPT\_Sample\_dataset2.csv"  
## [3] "ADAPT\_Sample\_dataset3.csv"

# Loading files in a single dataset

#Initializing dataframe  
df\_Data <- data.frame()  
  
#Reading multiple files  
for (i in filenames) {  
 filename <- paste0("./data/",i)  
 df <- read.csv(filename,stringsAsFactors = F)  
 df\_Data <- rbind(df\_Data,df)  
 rm(filename,df)  
}  
rm(filenames)  
  
#Reading Portcodes (.xlsx file)  
Port\_codes <- read\_xlsx(path="./data/Portcodes Dictionary.xlsx")

# Understanding the data

## Visualization of the dataset

View(df\_Data)

## Dimensions of the dataset

dim(df\_Data)

## [1] 49990 27

## Previewing some example data

head(df\_Data)

## ID CNTR\_N CNTR\_OPR\_C CNTR\_STATUS.C PURPOSE.C LENGTH\_Q CNTR\_TYPE\_C  
## 1 1 NYKU 9701818 ON F T 20 GP  
## 2 2 APHU 6602779 EG F T 40 GP  
## 3 3 PCIU 6007321 PI F T 40 RF  
## 4 4 HALU 5642133 HE F I 40 GP  
## 5 5 TEMU 1697343 YM F T 20 GP  
## 6 6 UACU 3529931 HL F T 20 GP  
## WT\_Q LOAD\_PORT\_C DISC\_PORT1\_C DEST\_PORT\_C DISC\_ABBR\_VESSEL\_M  
## 1 22080 MYPGU INCCU PAC BANDA  
## 2 22932 CNSHA IDPWG CMA TANCREDI  
## 3 15200 VNCLI AUMEL AUMEL   
## 4 27900 CNTXG AGAMEMNON  
## 5 15227 THBKK INMUN INMUN MATHU BHUM  
## 6 25000 MYPKG TRIST HONOLULU BR  
## DISC\_SERVICE\_C LOAD\_ABBR\_VESSEL\_M LOAD\_SERVICE\_C LOAD\_DT  
## 1 PGU DONG HAI CC1 2.018081e+13  
## 2 ASC MARCO P 118 IDO 2.018072e+13  
## 3 O KUALALUMPU AAA 2.018070e+13  
## 4 CIS NA  
## 5 TSB CONTI CHAMP CIS 2.018122e+13  
## 6 CIS HAMBURG E MD3 2.018093e+13  
## DISC\_DT GATE\_OUT\_DT GATE\_IN\_DT BATCH\_ID LOGISTICS\_PARTNER  
## 1 2.018080e+13 NA NA 201808 <NA>  
## 2 2.018072e+13 2.018072e+13 2.018072e+13 201807 <NA>  
## 3 NA NA NA 201807 <NA>  
## 4 2.018092e+13 2.018092e+13 NA 201809 <NA>  
## 5 2.018121e+13 NA NA 201812 <NA>  
## 6 2.018092e+13 NA NA 201809 <NA>  
## PRODUCT COMMODITY CARGO\_OWNER RF\_TEMP  
## 1 Weever Sea Food Renault NA  
## 2 Snail Sea Food Mitsubishi NA  
## 3 Snail Sea Food Mitsubishi 8  
## 4 Sorubim Sea Food Magna International NA  
## 5 pumpkin Vegetables Nissan NA  
## 6 delicata Vegetables Aviation Industry Corporation of China NA  
## DWELL\_DAYS HEIGHT\_Q  
## 1 3.1 9'6"  
## 2 7.6 9'6"  
## 3 4.3 9'6"  
## 4 2.7 9'6"  
## 5 2.8 9'6"  
## 6 9.6 9'6"

## Sampling 5 random rows

df\_Data[sample(1:nrow(df\_Data),5),]

## ID CNTR\_N CNTR\_OPR\_C CNTR\_STATUS.C PURPOSE.C LENGTH\_Q  
## 5161 5162 UACU 5088814 HL F T 40  
## 3754 3755 TEMU 6467180 CM F T 40  
## 5765 5766 CMAU 0063533 CM F T 20  
## 30810 30813 MRKU 9558846 MS F T 20  
## 23793 23796 CAIU 8684806 MD F T 40  
## CNTR\_TYPE\_C WT\_Q LOAD\_PORT\_C DISC\_PORT1\_C DEST\_PORT\_C  
## 5161 GP 25700 AUBNE THLCH   
## 3754 GP 31614 COCTG LKCMB INTUT  
## 5765 GP 19316 KRPUS LBBEY SYLTK  
## 30810 GP 16639 BEANR KRPUS   
## 23793 GP 26245 USORF IDBLW IDBLW  
## DISC\_ABBR\_VESSEL\_M DISC\_SERVICE\_C LOAD\_ABBR\_VESSEL\_M LOAD\_SERVICE\_C  
## 5161 BEA SCHULTE SAL   
## 3754 CMA MISSOURI PX2 CONTI PARIS NEN  
## 5765 CMA LAMARTIN ME5 COSCO SEINE ME3  
## 30810 MS HANOI LP3 MSC LA SPEZI AM4  
## 23793 MS SOFIA UE2   
## LOAD\_DT DISC\_DT GATE\_OUT\_DT GATE\_IN\_DT BATCH\_ID  
## 5161 NA 2.018113e+13 NA NA 201811  
## 3754 2.018083e+13 2.018082e+13 NA NA 201808  
## 5765 2.018112e+13 2.018111e+13 NA NA 201811  
## 30810 2.018122e+13 2.018122e+13 NA NA 201812  
## 23793 NA 2.018073e+13 NA NA 201807  
## LOGISTICS\_PARTNER PRODUCT COMMODITY  
## 5161 <NA> Star fruit Fruits  
## 3754 <NA> Snail Sea Food  
## 5765 <NA> Silverside Sea Food  
## 30810 <NA> Razor Sea Food  
## 23793 <NA> Shrimp Sea Food  
## CARGO\_OWNER RF\_TEMP DWELL\_DAYS HEIGHT\_Q  
## 5161 Boeing NA 2.8 9'6"  
## 3754 Mitsubishi NA 4.6 9'6"  
## 5765 Mitsui NA 8.2 8'6"  
## 30810 ThyssenKrupp NA 2.5 9'6"  
## 23793 Shandong Weiqiao Pioneering Group NA 2.0 8'6"

## Listing column names

names(df\_Data)

## [1] "ID" "CNTR\_N" "CNTR\_OPR\_C"   
## [4] "CNTR\_STATUS.C" "PURPOSE.C" "LENGTH\_Q"   
## [7] "CNTR\_TYPE\_C" "WT\_Q" "LOAD\_PORT\_C"   
## [10] "DISC\_PORT1\_C" "DEST\_PORT\_C" "DISC\_ABBR\_VESSEL\_M"  
## [13] "DISC\_SERVICE\_C" "LOAD\_ABBR\_VESSEL\_M" "LOAD\_SERVICE\_C"   
## [16] "LOAD\_DT" "DISC\_DT" "GATE\_OUT\_DT"   
## [19] "GATE\_IN\_DT" "BATCH\_ID" "LOGISTICS\_PARTNER"   
## [22] "PRODUCT" "COMMODITY" "CARGO\_OWNER"   
## [25] "RF\_TEMP" "DWELL\_DAYS" "HEIGHT\_Q"

## Listing just first 3 columns

head(names(df\_Data),3)

## [1] "ID" "CNTR\_N" "CNTR\_OPR\_C"

names(df\_Data[,1:3])

## [1] "ID" "CNTR\_N" "CNTR\_OPR\_C"

## Changing column names

orig\_names <- names(df\_Data)  
final\_names <- orig\_names  
final\_names[5] <- "PURPOSE\_C"  
final\_names[9] <- "POL"  
final\_names[10] <- "POD"  
final\_names[14] <- "LOAD\_VSL\_NAME"  
final\_names[12] <- "DISC\_VSL\_NAME"  
final\_names[4] <- "CONTR\_FOE"  
final\_names[6] <- "CNTR\_LENGTH"  
final\_names[8] <- "CNTR\_WGT\_KG"  
final\_names[27] <- "CNTR\_HEIGHT"  
  
names(df\_Data) <- final\_names  
print(final\_names)

## [1] "ID" "CNTR\_N" "CNTR\_OPR\_C"   
## [4] "CONTR\_FOE" "PURPOSE\_C" "CNTR\_LENGTH"   
## [7] "CNTR\_TYPE\_C" "CNTR\_WGT\_KG" "POL"   
## [10] "POD" "DEST\_PORT\_C" "DISC\_VSL\_NAME"   
## [13] "DISC\_SERVICE\_C" "LOAD\_VSL\_NAME" "LOAD\_SERVICE\_C"   
## [16] "LOAD\_DT" "DISC\_DT" "GATE\_OUT\_DT"   
## [19] "GATE\_IN\_DT" "BATCH\_ID" "LOGISTICS\_PARTNER"  
## [22] "PRODUCT" "COMMODITY" "CARGO\_OWNER"   
## [25] "RF\_TEMP" "DWELL\_DAYS" "CNTR\_HEIGHT"

## Formatting date columns

df\_Data$dt\_LOAD\_DT <- df\_Data$LOAD\_DT %>% as.character() %>% as.POSIXct(format="%Y%m%d%H%M%S")  
df\_Data$dt\_DISC\_DT <- df\_Data$DISC\_DT %>% as.character() %>% as.POSIXct(format="%Y%m%d%H%M%S")  
  
head(df\_Data[,c("dt\_LOAD\_DT","dt\_DISC\_DT")])

## dt\_LOAD\_DT dt\_DISC\_DT  
## 1 2018-08-08 04:25:30 2018-08-01 19:38:33  
## 2 2018-07-20 08:25:46 2018-07-15 05:06:40  
## 3 2018-07-01 06:48:28 <NA>  
## 4 <NA> 2018-09-21 08:18:27  
## 5 2018-12-20 01:49:42 2018-12-05 15:33:22  
## 6 2018-09-26 21:46:17 2018-09-18 17:45:33

## Obtaining data type of all columns

str(df\_Data)

## 'data.frame': 49990 obs. of 29 variables:  
## $ ID : int 1 2 3 4 5 6 7 8 9 10 ...  
## $ CNTR\_N : chr "NYKU 9701818" "APHU 6602779" "PCIU 6007321" "HALU 5642133" ...  
## $ CNTR\_OPR\_C : chr "ON" "EG" "PI" "HE" ...  
## $ CONTR\_FOE : chr "F" "F" "F" "F" ...  
## $ PURPOSE\_C : chr "T" "T" "T" "I" ...  
## $ CNTR\_LENGTH : int 20 40 40 40 20 20 20 40 20 40 ...  
## $ CNTR\_TYPE\_C : chr "GP" "GP" "RF" "GP" ...  
## $ CNTR\_WGT\_KG : int 22080 22932 15200 27900 15227 25000 17880 3880 26433 3660 ...  
## $ POL : chr "MYPGU" "CNSHA" "VNCLI" "CNTXG" ...  
## $ POD : chr "INCCU" "IDPWG" "AUMEL" "" ...  
## $ DEST\_PORT\_C : chr "" "" "AUMEL" "" ...  
## $ DISC\_VSL\_NAME : chr "PAC BANDA" "CMA TANCREDI" "" "AGAMEMNON" ...  
## $ DISC\_SERVICE\_C : chr "PGU" "ASC" "" "CIS" ...  
## $ LOAD\_VSL\_NAME : chr "DONG HAI" "MARCO P 118" "O KUALALUMPU" "" ...  
## $ LOAD\_SERVICE\_C : chr "CC1" "IDO" "AAA" "" ...  
## $ LOAD\_DT : num 2.02e+13 2.02e+13 2.02e+13 NA 2.02e+13 ...  
## $ DISC\_DT : num 2.02e+13 2.02e+13 NA 2.02e+13 2.02e+13 ...  
## $ GATE\_OUT\_DT : num NA 2.02e+13 NA 2.02e+13 NA ...  
## $ GATE\_IN\_DT : num NA 2.02e+13 NA NA NA ...  
## $ BATCH\_ID : int 201808 201807 201807 201809 201812 201809 201809 201810 201810 201807 ...  
## $ LOGISTICS\_PARTNER: chr NA NA NA NA ...  
## $ PRODUCT : chr "Weever" "Snail" "Snail" "Sorubim" ...  
## $ COMMODITY : chr "Sea Food" "Sea Food" "Sea Food" "Sea Food" ...  
## $ CARGO\_OWNER : chr "Renault" "Mitsubishi" "Mitsubishi" "Magna International" ...  
## $ RF\_TEMP : int NA NA 8 NA NA NA NA NA NA NA ...  
## $ DWELL\_DAYS : num 3.1 7.6 4.3 2.7 2.8 9.6 3.5 12.1 6.6 9.6 ...  
## $ CNTR\_HEIGHT : chr "9'6\"" "9'6\"" "9'6\"" "9'6\"" ...  
## $ dt\_LOAD\_DT : POSIXct, format: "2018-08-08 04:25:30" "2018-07-20 08:25:46" ...  
## $ dt\_DISC\_DT : POSIXct, format: "2018-08-01 19:38:33" "2018-07-15 05:06:40" ...

## Obtaining a quick summary of each column

summary(df\_Data)

## ID CNTR\_N CNTR\_OPR\_C CONTR\_FOE   
## Min. : 1 Length:49990 Length:49990 Length:49990   
## 1st Qu.:12499 Class :character Class :character Class :character   
## Median :24999 Mode :character Mode :character Mode :character   
## Mean :24998   
## 3rd Qu.:37496   
## Max. :49993   
##   
## PURPOSE\_C CNTR\_LENGTH CNTR\_TYPE\_C CNTR\_WGT\_KG   
## Length:49990 Min. :20.00 Length:49990 Min. : 1730   
## Class :character 1st Qu.:20.00 Class :character 1st Qu.: 9400   
## Mode :character Median :40.00 Mode :character Median :21154   
## Mean :30.62 Mean :18407   
## 3rd Qu.:40.00 3rd Qu.:27223   
## Max. :45.00 Max. :37900   
##   
## POL POD DEST\_PORT\_C   
## Length:49990 Length:49990 Length:49990   
## Class :character Class :character Class :character   
## Mode :character Mode :character Mode :character   
##   
##   
##   
##   
## DISC\_VSL\_NAME DISC\_SERVICE\_C LOAD\_VSL\_NAME   
## Length:49990 Length:49990 Length:49990   
## Class :character Class :character Class :character   
## Mode :character Mode :character Mode :character   
##   
##   
##   
##   
## LOAD\_SERVICE\_C LOAD\_DT DISC\_DT   
## Length:49990 Min. :2.018e+13 Min. :2.018e+13   
## Class :character 1st Qu.:2.018e+13 1st Qu.:2.018e+13   
## Mode :character Median :2.018e+13 Median :2.018e+13   
## Mean :2.018e+13 Mean :2.018e+13   
## 3rd Qu.:2.018e+13 3rd Qu.:2.018e+13   
## Max. :2.018e+13 Max. :2.018e+13   
## NA's :9472 NA's :9922   
## GATE\_OUT\_DT GATE\_IN\_DT BATCH\_ID   
## Min. :2.018e+13 Min. :2.018e+13 Min. :201807   
## 1st Qu.:2.018e+13 1st Qu.:2.018e+13 1st Qu.:201808   
## Median :2.018e+13 Median :2.018e+13 Median :201809   
## Mean :2.018e+13 Mean :2.018e+13 Mean :201809   
## 3rd Qu.:2.018e+13 3rd Qu.:2.018e+13 3rd Qu.:201811   
## Max. :2.018e+13 Max. :2.018e+13 Max. :201812   
## NA's :40533 NA's :40684   
## LOGISTICS\_PARTNER PRODUCT COMMODITY   
## Length:49990 Length:49990 Length:49990   
## Class :character Class :character Class :character   
## Mode :character Mode :character Mode :character   
##   
##   
##   
##   
## CARGO\_OWNER RF\_TEMP DWELL\_DAYS CNTR\_HEIGHT   
## Length:49990 Min. :-30.00 Min. : 2.000 Length:49990   
## Class :character 1st Qu.:-20.00 1st Qu.: 4.200 Class :character   
## Mode :character Median :-10.00 Median : 6.400 Mode :character   
## Mean :-10.23 Mean : 6.392   
## 3rd Qu.: 0.00 3rd Qu.: 8.400   
## Max. : 10.00 Max. :15.000   
## NA's :46969   
## dt\_LOAD\_DT dt\_DISC\_DT   
## Min. :2018-06-29 18:35:53 Min. :2018-07-01 00:00:19   
## 1st Qu.:2018-08-15 08:25:12 1st Qu.:2018-08-15 08:36:00   
## Median :2018-09-30 10:42:39 Median :2018-09-30 11:49:51   
## Mean :2018-09-30 10:02:55 Mean :2018-09-30 19:47:11   
## 3rd Qu.:2018-11-15 06:44:13 3rd Qu.:2018-11-16 02:24:00   
## Max. :2018-12-31 21:25:23 Max. :2018-12-31 23:56:44   
## NA's :9475 NA's :9922

## Determine number of NAs in a single column

sum(is.na(df\_Data$RF\_TEMP))

## [1] 46969

## Determine number of NAs in each column of the entire dataset

colSums(is.na(df\_Data))

## ID CNTR\_N CNTR\_OPR\_C CONTR\_FOE   
## 0 0 0 0   
## PURPOSE\_C CNTR\_LENGTH CNTR\_TYPE\_C CNTR\_WGT\_KG   
## 0 0 0 0   
## POL POD DEST\_PORT\_C DISC\_VSL\_NAME   
## 0 0 0 0   
## DISC\_SERVICE\_C LOAD\_VSL\_NAME LOAD\_SERVICE\_C LOAD\_DT   
## 0 0 0 9472   
## DISC\_DT GATE\_OUT\_DT GATE\_IN\_DT BATCH\_ID   
## 9922 40533 40684 0   
## LOGISTICS\_PARTNER PRODUCT COMMODITY CARGO\_OWNER   
## 46779 13 13 13   
## RF\_TEMP DWELL\_DAYS CNTR\_HEIGHT dt\_LOAD\_DT   
## 46969 0 0 9475   
## dt\_DISC\_DT   
## 9922

## Determine % of NAs in each column of the entire dataset

round(100\*colSums(is.na(df\_Data)) / nrow(df\_Data),0)

## ID CNTR\_N CNTR\_OPR\_C CONTR\_FOE   
## 0 0 0 0   
## PURPOSE\_C CNTR\_LENGTH CNTR\_TYPE\_C CNTR\_WGT\_KG   
## 0 0 0 0   
## POL POD DEST\_PORT\_C DISC\_VSL\_NAME   
## 0 0 0 0   
## DISC\_SERVICE\_C LOAD\_VSL\_NAME LOAD\_SERVICE\_C LOAD\_DT   
## 0 0 0 19   
## DISC\_DT GATE\_OUT\_DT GATE\_IN\_DT BATCH\_ID   
## 20 81 81 0   
## LOGISTICS\_PARTNER PRODUCT COMMODITY CARGO\_OWNER   
## 94 0 0 0   
## RF\_TEMP DWELL\_DAYS CNTR\_HEIGHT dt\_LOAD\_DT   
## 94 0 0 19   
## dt\_DISC\_DT   
## 20

## Summarizing information

df\_Data %>% group\_by(CONTR\_FOE) %>% summarise(Weight\_avg=mean(CNTR\_WGT\_KG))

## # A tibble: 2 x 2  
## CONTR\_FOE Weight\_avg  
## <chr> <dbl>  
## 1 E 3177.  
## 2 F 21074.