lab10_stat123

Koki Itagaki

2023-03-28

1. Install the following packages: "dplyr", "ggplot2", "corrplot", "car", "lmtest", and "caret". Then load the dataset "stock-Canada.csv". You can find a description of the dataset in the file "stock-dictionary.csv". Your task is to predict the value of the stock. Please explain the output, your solutions, and the given codes.

```
library("dplyr")

##
## 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")
dairy_sttocks<-read.csv(file ="/Users/itagakikouki/stat123/lab10/stock-
Canada.csv")
dim(dairy_sttocks)

## [1] 38428    16

head(dairy_sttocks)

##   REF_DATE                GEO          DGUID       Stocks       Commodity
## 1  1970-01             Canada 2016A000011124 Total stocks Creamery butter
## 2  1970-01             Canada 2016A000011124 Total stocks  Cheddar cheese
## 3  1970-01             Canada 2016A000011124 Total stocks  Variety cheese
## 4  1970-01             Canada 2016A000011124 Total stocks     Whey butter
## 5  1970-01             Canada 2016A000011124 Total stocks  Process cheese
## 6  1970-01 Maritime provinces                Total stocks  Cheddar cheese
##      UOM UOM_ID SCALAR_FACTOR SCALAR_ID  VECTOR COORDINATE VALUE STATUS
SYMBOL
## 1 Tonnes    287         units         0 v382775      1.1.1 40829
NA
## 2 Tonnes    287         units         0 v382812      1.1.3 36681
NA
## 3 Tonnes    287         units         0 v382827      1.1.4  2537
NA
## 4 Tonnes    287         units         0 v382840      1.1.5   116
NA
```

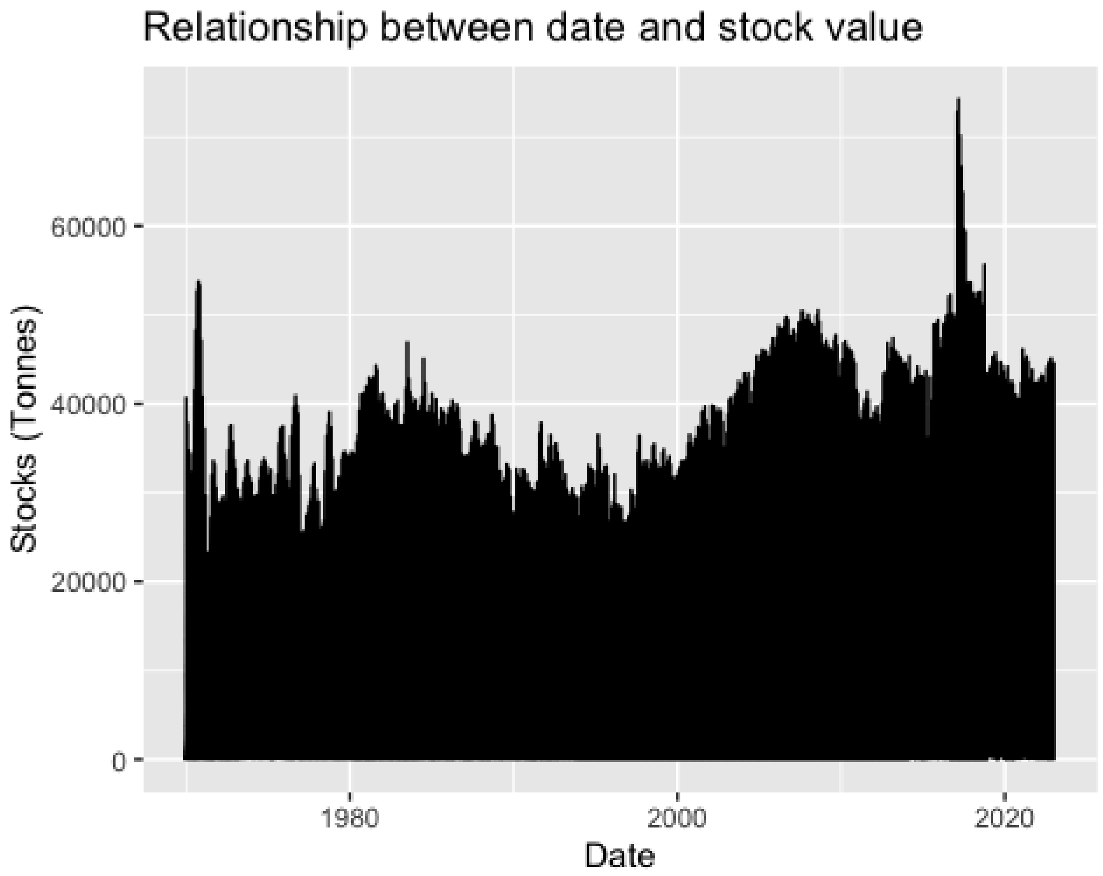
```
## 5 Tonnes    287        units          0 v382850      1.1.6  3021
NA
## 6 Tonnes    287        units          0 v382813      3.1.3   326
NA
##   TERMINATED DECIMALS
## 1                   0
## 2                   0
## 3                   0
## 4                   0
## 5                   0
## 6                   0
```

```
#(a) Clean the data by removing null and missing values. You may also choose
to remove columns that have a high number of missing values.
#(b) Convert the "REF_DATE" column to a date format using the "mutate()"
function as follows:mutate(REF_DATE = as.Date(paste0(REF_DATE, "-01"), format
= "%Y-%m-%d"))
#Note: This code uses the dplyr package's "mutate()" function to create a new
column called
#"REF_DATE" in a dataframe, where "REF_DATE" is an existing column in the
same dataframe.
#The purpose of this code is to convert the "REF_DATE" column, which is
currently in a characterformat, into a date format. The "as.Date()" function
is used to convert the character strings in the "REF_DATE" column into
#date format. The "paste0()" function is used to concatenate the year-month
values in the
#"REF_DATE" column with "-01", which represents the first day of the month,
creating a new
#string in the format "YYYY-MM-01". This new string is then passed to the
"as.Date()" function as the first argument, which converts it to a date
format.
#The "format" argument in the "as.Date()" function is used to specify the
format of the input string. In this case, the format is "%Y-%m-%d", which
indicates that the input string is in the format"YYYY-MM-DD". Since the input
string only contains the year and month, "-01" is added to represent the
first day of the month.

dairy_sttocks<- dairy_sttocks%>%
  select(REF_DATE,GEO,Stocks,Commodity,VALUE)%>%
  mutate(REF_DATE = as.Date(paste0(REF_DATE,"-01"), format = "%Y-%m-%d"))
dim(dairy_sttocks)
```

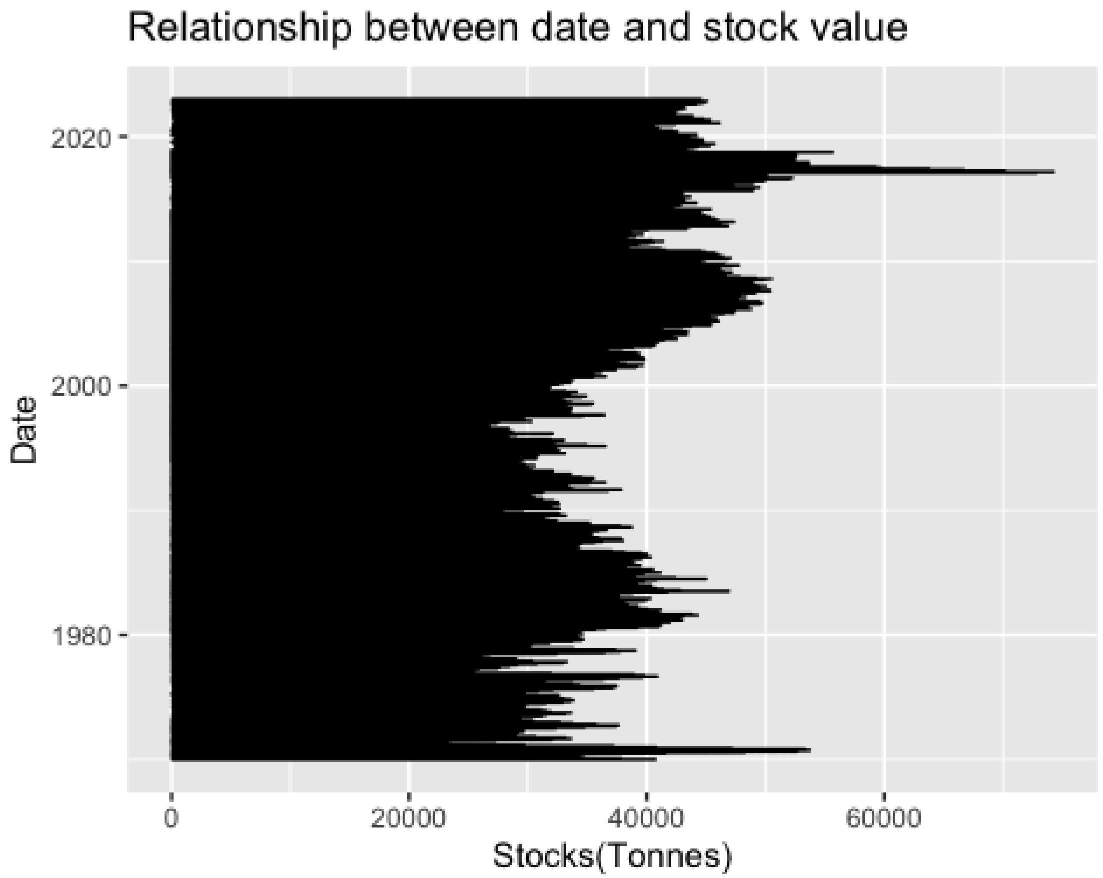
```
## [1] 38428     5
```

```
head(dairy_sttocks)
```

```
##     REF_DATE                GEO       Stocks       Commodity VALUE
## 1 1970-01-01             Canada Total stocks Creamery butter 40829
## 2 1970-01-01             Canada Total stocks  Cheddar cheese 36681
## 3 1970-01-01             Canada Total stocks  Variety cheese  2537
## 4 1970-01-01             Canada Total stocks     Whey butter   116
## 5 1970-01-01             Canada Total stocks  Process cheese  3021
## 6 1970-01-01 Maritime provinces Total stocks  Cheddar cheese   326

#SO many NA's
summary(dairy_sttocks$VALUE)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's 
##       0     530    2250    7007    9444   74242   11928

#Since NA exists
sd(dairy_sttocks$VALUE)

## [1] NA

#remove na
dairy_sttocks<-na.omit(dairy_sttocks)



#(c) Process the data using the "summary()" function and calculate the
standard deviation for the column that has numeric values.
summary(dairy_sttocks$VALUE)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##       0     530    2250    7007    9444   74242

sd(dairy_sttocks$VALUE)

## [1] 9970.362

#(d) Use the "ggplot()" function to visualize the relationship between each
column and the value of the stock.
ggplot(data = dairy_sttocks, aes(x = REF_DATE, y = VALUE)) + geom_line() +
labs(x = "Date", y = "Stocks (Tonnes)", title = "Relationship between date
and stock value")
```



```
ggplot(dairy_sttocks, mapping = aes(x = REF_DATE, y = VALUE)) + geom_line()
+labs(x = "Date", y = "Stocks(Tonnes)", title = "Relationship between date
and stock value") +
  coord_flip()
```



```
#(e) Split the data into 80% train and 20% test using the following code:
# Split the data into training and test sets
set.seed(2023)
 train_index<- sample(1:nrow(dairy_sttocks),size =
round(0.8*nrow(dairy_sttocks)))
 train_data<-dairy_sttocks[train_index, ]
 test_data<- dairy_sttocks[-train_index,]

#set.seed(123)
#train_index <- sample(1:nrow(dairy_stocks), size = round(0.8 *
nrow(dairy_stocks)))
#train_data <- dairy_stocks[train_index,]
#test_data <- dairy_stocks[-train_index,]
#Note: To split the data, you can use the sample() function and specify the
percentage of data in each split. For example, "size = round(0.8 *
nrow(dairy_stocks))" means 80% of the data will be used for training, and the
rest will be used for testing.
#(f) Build a multiple linear regression model to predict the value of the
stock. The data type of some columns is not numerical (they are categorical).
It is acceptable to consider those columns in the model, but in a data
science project, you need to consider the data type.
 model1<-lm(VALUE~REF_DATE +GEO + Stocks + Commodity, data = train_data)
 model1
```

```
##
## Call:
## lm(formula = VALUE ~ REF_DATE + GEO + Stocks + Commodity, data =
train_data)
##
## Coefficients:
##                                (Intercept)
##                                  3.050e+03
##                                   REF_DATE
##                                  2.647e-01
##                      GEOAtlantic provinces
##                                 -1.085e+02
##                        GEOBritish Columbia
##                                 -8.662e+02
##                                  GEOCanada
##                                  2.033e+04
##                                GEOManitoba
##                                 -1.622e+03
##                      GEOMaritime provinces
##                                  6.843e+02
##                           GEONew Brunswick
##                                  1.350e+03
##                             GEONova Scotia
##                                  1.464e+03
##                                 GEOOntario
##                                  9.112e+03
##                         GEOOther Provinces
##                                  3.381e+03
##                    GEOPrince Edward Island
##                                  1.379e+03
##                                  GEOQuebec
##                                  8.478e+03
##                            GEOSaskatchewan
##                                 -6.106e+02
##          StocksRetail and wholesale stocks
##                                 -1.109e+04
##                         StocksTotal stocks
##                                  7.237e+02
##  CommodityConcentrated partly skimmed milk
##                                 -2.628e+04
##            CommodityConcentrated skim milk
##                                 -2.668e+04
##           CommodityConcentrated whole milk
##                                 -2.263e+04
##                    CommodityCondensed milk
##                                 -2.494e+04
##               CommodityCondensed skim milk
##                                 -2.549e+04
##                   CommodityCreamery butter
##                                 -5.946e+03
##                   CommodityEvaporated milk
##                                 -4.722e+03
##              CommodityEvaporated skim milk
##                                 -2.537e+04
## CommodityPartly skimmed evaporated milk 2%
##                                 -2.443e+04
##               CommodityPowdered buttermilk
##                                 -2.648e+04
##                    CommodityProcess cheese
##                                 -1.773e+04
##                  CommoditySkim milk powder
##                                 -2.217e+03
##  CommoditySweetened concentrated skim milk
##                                 -2.687e+04
## CommoditySweetened concentrated whole milk
##                                 -2.523e+04
##                    CommodityVariety cheese
##                                 -7.622e+03
##                       CommodityWhey butter
##                                 -2.597e+04
##                       CommodityWhey powder
##                                 -1.993e+04
##                 CommodityWhole milk powder
##                                 -2.595e+04
```

```
summary(model1)
```

```
##
## Call:
## lm(formula = VALUE ~ REF_DATE + GEO + Stocks + Commodity, data =
train_data)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -19706  -3184   -322   2223  48519
##
## Coefficients:
##                                              Estimate Std. Error  t value
## (Intercept)                                 3.050e+03  2.264e+02   13.474
## REF_DATE                                    2.647e-01  9.191e-03   28.804
## GEOAtlantic provinces                      -1.085e+02  3.662e+02   -0.296
## GEOBritish Columbia                        -8.662e+02  2.454e+02   -3.530
## GEOCanada                                   2.033e+04  1.845e+02  110.177
## GEOManitoba                                -1.622e+03  2.252e+02   -7.205
## GEOMaritime provinces                       6.843e+02  3.527e+02    1.940
## GEONew Brunswick                            1.350e+03  4.421e+02    3.053
## GEONova Scotia                              1.464e+03  4.608e+02    3.177
## GEOOntario                                  9.112e+03  1.867e+02   48.802
## GEOOther Provinces                          3.381e+03  2.278e+02   14.842
## GEOPrince Edward Island                     1.379e+03  4.607e+02    2.993
## GEOQuebec                                   8.478e+03  1.919e+02   44.183
## GEOSaskatchewan                            -6.106e+02  2.605e+02   -2.344
## StocksRetail and wholesale stocks          -1.109e+04  1.458e+02  -76.095
## StocksTotal stocks                          7.237e+02  1.064e+02    6.802
## CommodityConcentrated partly skimmed milk  -2.628e+04  2.832e+02  -92.793
## CommodityConcentrated skim milk            -2.668e+04  2.832e+02  -94.186
## CommodityConcentrated whole milk           -2.263e+04  2.699e+02  -83.865
## CommodityCondensed milk                    -2.494e+04  1.010e+03  -24.698
## CommodityCondensed skim milk               -2.549e+04  1.042e+03  -24.458
## CommodityCreamery butter                   -5.946e+03  1.127e+02  -52.775
## CommodityEvaporated milk                   -4.722e+03  1.098e+03   -4.302
## CommodityEvaporated skim milk              -2.537e+04  9.800e+02  -25.883
## CommodityPartly skimmed evaporated milk 2% -2.443e+04  1.098e+03  -22.260
## CommodityPowdered buttermilk               -2.648e+04  2.548e+02 -103.933
## CommodityProcess cheese                    -1.773e+04  2.127e+02  -83.344
## CommoditySkim milk powder                  -2.217e+03  2.611e+02   -8.492
## CommoditySweetened concentrated skim milk  -2.687e+04  3.513e+02  -76.493
## CommoditySweetened concentrated whole milk -2.523e+04  6.282e+02  -40.168
## CommodityVariety cheese                    -7.622e+03  1.301e+02  -58.593
## CommodityWhey butter                       -2.597e+04  2.740e+02  -94.753
## CommodityWhey powder                       -1.993e+04  2.558e+02  -77.916
## CommodityWhole milk powder                 -2.595e+04  2.855e+02  -90.899
##                                            Pr(>|t|)    
## (Intercept)                                 < 2e-16 ***
## REF_DATE                                    < 2e-16 ***
## GEOAtlantic provinces                      0.766957    
## GEOBritish Columbia                        0.000416 ***
## GEOCanada                                   < 2e-16 ***
## GEOManitoba                                6.00e-13 ***
## GEOMaritime provinces                      0.052373 .  
## GEONew Brunswick                           0.002265 ** 
## GEONova Scotia                             0.001488 ** 
## GEOOntario                                  < 2e-16 ***
## GEOOther Provinces                          < 2e-16 ***
## GEOPrince Edward Island                    0.002769 ** 
## GEOQuebec                                   < 2e-16 ***
## GEOSaskatchewan                            0.019069 *  
## StocksRetail and wholesale stocks           < 2e-16 ***
## StocksTotal stocks                         1.06e-11 ***
## CommodityConcentrated partly skimmed milk   < 2e-16 ***
## CommodityConcentrated skim milk             < 2e-16 ***
## CommodityConcentrated whole milk            < 2e-16 ***
## CommodityCondensed milk                     < 2e-16 ***
## CommodityCondensed skim milk                < 2e-16 ***
## CommodityCreamery butter                    < 2e-16 ***
## CommodityEvaporated milk                   1.70e-05 ***
## CommodityEvaporated skim milk               < 2e-16 ***
## CommodityPartly skimmed evaporated milk 2%  < 2e-16 ***
## CommodityPowdered buttermilk                < 2e-16 ***
## CommodityProcess cheese                     < 2e-16 ***
## CommoditySkim milk powder                   < 2e-16 ***
## CommoditySweetened concentrated skim milk   < 2e-16 ***
## CommoditySweetened concentrated whole milk  < 2e-16 ***
## CommodityVariety cheese                     < 2e-16 ***
## CommodityWhey butter                        < 2e-16 ***
## CommodityWhey powder                        < 2e-16 ***
## CommodityWhole milk powder                  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5660 on 21166 degrees of freedom
## Multiple R-squared:  0.6828, Adjusted R-squared:  0.6823
## F-statistic:  1380 on 33 and 21166 DF,  p-value: < 2.2e-16
```

```
#(g) Identify variables that should be removed from the model.
 cat("The p value of GEOAtlantic provinces is 0.766957 which is higher than
     a = 0.05. So I remove it")
```

```
## The p value of GEOAtlantic provinces is 0.766957 which is higher than
##      a = 0.05. So I remove it
```