

Assignment3

91222033 ZHANG XINYUE

2022-11-04

Assignment3

This assignment will clean up and basically summarize the data given.

```
##### read the data####
```

```
dataset <- read.csv("C:/Users/78209/OneDrive/ /price_index_Feb20201 (1).csv",header=TRUE)
```

```
#####Just look at the first five pieces of data###
```

```
head(dataset, n=5)
```

```
##          SER_REF TIME_REF DATA_VAL STATUS UNITS Subject
## 1 CPIM.SE1041F  2006.11      1000  FINAL Index      CPI
## 2 CPIM.SE1041F  2006.12        995  FINAL Index      CPI
## 3 CPIM.SE1041F  2007.01      1007  FINAL Index      CPI
## 4 CPIM.SE1041F  2007.02      1012  FINAL Index      CPI
## 5 CPIM.SE1041F  2007.03      1020  FINAL Index      CPI
##                                     Group Series_title_1      Series_title_2
## 1 CPI Monthly Rents (Broad Regions) Auckland Actual rentals for housing
## 2 CPI Monthly Rents (Broad Regions) Auckland Actual rentals for housing
## 3 CPI Monthly Rents (Broad Regions) Auckland Actual rentals for housing
## 4 CPI Monthly Rents (Broad Regions) Auckland Actual rentals for housing
## 5 CPI Monthly Rents (Broad Regions) Auckland Actual rentals for housing
## Series_title_3
## 1          Flow
## 2          Flow
## 3          Flow
## 4          Flow
## 5          Flow
```

```
#####Observe the overall properties of the dataset###
```

```
summary(dataset)
```

```
##          SER_REF          TIME_REF          DATA_VAL          STATUS
## Length:1203      Min.   :2006      Min.   : 969      Length:1203
## Class :character 1st Qu.:2010      1st Qu.:1108      Class :character
## Mode  :character Median :2014      Median :1216      Mode  :character
##                Mean   :2014      Mean   :1253
##                3rd Qu.:2017      3rd Qu.:1392
##                Max.   :2021      Max.   :1683
```

```
##      UNITS          Subject          Group          Series_title_1
## Length:1203      Length:1203      Length:1203      Length:1203
## Class :character  Class :character  Class :character  Class :character
## Mode  :character  Mode  :character  Mode  :character  Mode  :character
##
##
##
## Series_title_2      Series_title_3
## Length:1203      Length:1203
## Class :character  Class :character
## Mode  :character  Mode  :character
##
##
##
```

```
#####View detailed data types and other information##
str(dataset)
```

```
## 'data.frame':    1203 obs. of  10 variables:
## $ SER_REF       : chr  "CPIM.SE1041F" "CPIM.SE1041F" "CPIM.SE1041F" "CPIM.SE1041F" ...
## $ TIME_REF      : num  2006 2006 2007 2007 2007 ...
## $ DATA_VAL     : int  1000 995 1007 1012 1020 1023 1045 1033 1037 1058 ...
## $ STATUS        : chr  "FINAL" "FINAL" "FINAL" "FINAL" ...
## $ UNITS         : chr  "Index" "Index" "Index" "Index" ...
## $ Subject       : chr  "CPI" "CPI" "CPI" "CPI" ...
## $ Group         : chr  "CPI Monthly Rents (Broad Regions)" "CPI Monthly Rents (Broad Regions)" "CPI
## $ Series_title_1: chr  "Auckland" "Auckland" "Auckland" "Auckland" ...
## $ Series_title_2: chr  "Actual rentals for housing" "Actual rentals for housing" "Actual rentals for
## $ Series_title_3: chr  "Flow" "Flow" "Flow" "Flow" ...
```

```
#####Check the price variable for missing values####
sum(is.na(dataset$DATA_VAL))
```

```
## [1] 0
```

```
#####Check all columns for missing values####
colSums(is.na(dataset))
```

```
##      SER_REF      TIME_REF      DATA_VAL      STATUS      UNITS
##          0          0          0          0          0
##      Subject      Group Series_title_1 Series_title_2 Series_title_3
##          0          0          0          0          0
```

```
#####Calculate the size of a data set####
object.size(dataset)
```

```
## 94776 bytes
```

```
print(object.size(dataset),units="Mb")
```

```
## 0.1 Mb
```

```
#####Create a data table with only SER_REF and DATA_VAL columns#####
classes <- sapply(dataset, class)
classes[c(-1,-3)] <- rep("NULL", length(classes)-2)
data <- read.csv("C:/Users/78209/OneDrive/ /price_index_Feb20201 (1).csv", colClasses = classes)
```

```
#####The number of title1 data in the Rest of North Island was counted###
table(dataset$Series_title_1 %in% c('Rest of North Island'))
```

```
##
## FALSE TRUE
## 1031 172
```

```
#####Use the Freq function in the DescTools package to generate a frequency distribution table about v.
library(DescTools)
```

```
## Warning: 'DescTools' R 4.2.2
```

```
price <- Freq(dataset$DATA_VAL)
price
```

	level	freq	perc	cumfreq	cumperc
## 1	[950,1e+03]	12	1.0%	12	1.0%
## 2	(1e+03,1.05e+03]	71	5.9%	83	6.9%
## 3	(1.05e+03,1.1e+03]	195	16.2%	278	23.1%
## 4	(1.1e+03,1.15e+03]	162	13.5%	440	36.6%
## 5	(1.15e+03,1.2e+03]	114	9.5%	554	46.1%
## 6	(1.2e+03,1.25e+03]	113	9.4%	667	55.4%
## 7	(1.25e+03,1.3e+03]	76	6.3%	743	61.8%
## 8	(1.3e+03,1.35e+03]	72	6.0%	815	67.7%
## 9	(1.35e+03,1.4e+03]	109	9.1%	924	76.8%
## 10	(1.4e+03,1.45e+03]	93	7.7%	1'017	84.5%
## 11	(1.45e+03,1.5e+03]	68	5.7%	1'085	90.2%
## 12	(1.5e+03,1.55e+03]	84	7.0%	1'169	97.2%
## 13	(1.55e+03,1.6e+03]	17	1.4%	1'186	98.6%
## 14	(1.6e+03,1.65e+03]	12	1.0%	1'198	99.6%
## 15	(1.65e+03,1.7e+03]	5	0.4%	1'203	100.0%

```
#####Just look at Series titles 1,2,3,To observe the specific information and state of the product##
x <- as.matrix(dataset[, 8:10])
x
```

	Series_title_1	Series_title_2	Series_title_3
## [1,]	"Auckland"	"Actual rentals for housing"	"Flow"
## [2,]	"Auckland"	"Actual rentals for housing"	"Flow"
## [3,]	"Auckland"	"Actual rentals for housing"	"Flow"
## [4,]	"Auckland"	"Actual rentals for housing"	"Flow"
## [5,]	"Auckland"	"Actual rentals for housing"	"Flow"
## [6,]	"Auckland"	"Actual rentals for housing"	"Flow"
## [7,]	"Auckland"	"Actual rentals for housing"	"Flow"
## [8,]	"Auckland"	"Actual rentals for housing"	"Flow"
## [9,]	"Auckland"	"Actual rentals for housing"	"Flow"

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

```
## [1198,] "National"          "Actual rentals for housing" "Stock"
## [1199,] "National"          "Actual rentals for housing" "Stock"
## [1200,] "National"          "Actual rentals for housing" "Stock"
## [1201,] "National"          "Actual rentals for housing" "Stock"
## [1202,] "National"          "Actual rentals for housing" "Stock"
## [1203,] "National"          "Actual rentals for housing" "Stock"
```

```
##### Use the factor function for column "Series_title_1"
```

```
area<-factor(c('Auckland','Wellington','Rest of North Island','Canterbury','Rest of South Island','National'))
class(area)
```

```
## [1] "factor"
```

```
levels(area)
```

```
## [1] "Auckland"          "Canterbury"          "National"
## [4] "Rest of North Island" "Rest of South Island" "Wellington"
```

```
nlevels(area)
```

```
## [1] 6
```

```
#1.reading data
```

```
data<- read.csv("C:/Users/78209/OneDrive/ /price_index_Feb20201 (1).csv",header = T)
colnames(data)
```

```
## [1] "SER_REF"          "TIME_REF"          "DATA_VAL"          "STATUS"
## [5] "UNITS"            "Subject"           "Group"             "Series_title_1"
## [9] "Series_title_2"   "Series_title_3"
```

```
table(data$SER_REF) #There are seven categories of items
```

```
##
## CPIM.SE1041F CPIM.SE2041F CPIM.SE3041F CPIM.SE5041F CPIM.SE6041F CPIM.SE9041F
##          172          172          172          172          172          172
## CPIM.SE9041S
##          171
```

```
str(data) #View the properties of the variable
```

```
## 'data.frame':    1203 obs. of  10 variables:
## $ SER_REF       : chr  "CPIM.SE1041F" "CPIM.SE1041F" "CPIM.SE1041F" "CPIM.SE1041F" ...
## $ TIME_REF      : num  2006 2006 2007 2007 2007 ...
## $ DATA_VAL     : int   1000 995 1007 1012 1020 1023 1045 1033 1037 1058 ...
## $ STATUS        : chr   "FINAL" "FINAL" "FINAL" "FINAL" ...
## $ UNITS         : chr   "Index" "Index" "Index" "Index" ...
## $ Subject       : chr   "CPI" "CPI" "CPI" "CPI" ...
## $ Group         : chr   "CPI Monthly Rents (Broad Regions)" "CPI Monthly Rents (Broad Regions)" "CPI
## $ Series_title_1: chr   "Auckland" "Auckland" "Auckland" "Auckland" ...
## $ Series_title_2: chr   "Actual rentals for housing" "Actual rentals for housing" "Actual rentals for
## $ Series_title_3: chr   "Flow" "Flow" "Flow" "Flow" ...
```

```
data$DATA_VAL<-as.numeric(data$DATA_VAL)
class(data$DATA_VAL)
```

```
## [1] "numeric"
```

```
#Calculate the average of the prices of seven categories of items
x<-list(a=data$DATA_VAL[which(data$SER_REF=="CPIM.SE1041F")],
        b=data$DATA_VAL[which(data$SER_REF=="CPIM.SE2041F")],
        c=data$DATA_VAL[which(data$SER_REF=="CPIM.SE3041F")],
        d=data$DATA_VAL[which(data$SER_REF=="CPIM.SE5041F")],
        e=data$DATA_VAL[which(data$SER_REF=="CPIM.SE6041F")],
        f=data$DATA_VAL[which(data$SER_REF=="CPIM.SE9041F")],
        g=data$DATA_VAL[which(data$SER_REF=="CPIM.SE9041S")])
#Construct a matrix of X with the categories and values corresponding to X
#Take the mean of the elements in x
sapply(x,mean)
```

```
##          a          b          c          d          e          f          g
## 1279.773 1263.564 1234.849 1262.471 1221.855 1258.099 1250.930
```