Obembe_wk5_assignment

October 3, 2021

0.0.1 Reading in the dataset

[2]:
```
require(ggplot2)
```

```
Loading required package: ggplot2
```

[3]:
```
data(diamonds)
head(diamonds)
```

carat	cut	color	clarity	depth	table	price	x	y	z
0.23	Ideal	E	SI2	61.5	55	326	3.95	3.98	2.43
0.21	Premium	E	SI1	59.8	61	326	3.89	3.84	2.31
0.23	Good	E	VS1	56.9	65	327	4.05	4.07	2.31
0.29	Premium	I	VS2	62.4	58	334	4.20	4.23	2.63
0.31	Good	J	SI2	63.3	58	335	4.34	4.35	2.75
0.24	Very Good	J	VVS2	62.8	57	336	3.94	3.96	2.48

0.0.2 Exploring filtering function

[4]:
```
# Filter the table for 'cut' = Premium

library("dplyr")
```

```
Warning message:
"package 'dplyr' was built under R version 3.6.3"
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
```

[5]:
```
df <- diamonds
head(df)
```

carat	cut	color	clarity	depth	table	price	x	y	z
0.23	Ideal	E	SI2	61.5	55	326	3.95	3.98	2.43
0.21	Premium	E	SI1	59.8	61	326	3.89	3.84	2.31
0.23	Good	E	VS1	56.9	65	327	4.05	4.07	2.31
0.29	Premium	I	VS2	62.4	58	334	4.20	4.23	2.63
0.31	Good	J	SI2	63.3	58	335	4.34	4.35	2.75
0.24	Very Good	J	VVS2	62.8	57	336	3.94	3.96	2.48

[7]:
```
prem<- filter(df, cut == "Premium")
head(prem)
```

carat	cut	color	clarity	depth	table	price	x	y	z
0.21	Premium	E	SI1	59.8	61	326	3.89	3.84	2.31
0.29	Premium	I	VS2	62.4	58	334	4.20	4.23	2.63
0.22	Premium	F	SI1	60.4	61	342	3.88	3.84	2.33
0.20	Premium	E	SI2	60.2	62	345	3.79	3.75	2.27
0.32	Premium	E	I1	60.9	58	345	4.38	4.42	2.68
0.24	Premium	I	VS1	62.5	57	355	3.97	3.94	2.47

[8]:
```
# Filter for color = 'J'

colr <- filter(df,color=="J")
head(colr)
```

carat	cut	color	clarity	depth	table	price	x	y	z
0.31	Good	J	SI2	63.3	58	335	4.34	4.35	2.75
0.24	Very Good	J	VVS2	62.8	57	336	3.94	3.96	2.48
0.30	Good	J	SI1	64.0	55	339	4.25	4.28	2.73
0.23	Ideal	J	VS1	62.8	56	340	3.93	3.90	2.46
0.31	Ideal	J	SI2	62.2	54	344	4.35	4.37	2.71
0.30	Good	J	SI1	63.4	54	351	4.23	4.29	2.70

0.0.3 Using Select Function

[10]:
```
df1 = df %>% select(carat,cut,color, price)
head(df1)
```

carat	cut	color	price
0.23	Ideal	E	326
0.21	Premium	E	326
0.23	Good	E	327
0.29	Premium	I	334
0.31	Good	J	335
0.24	Very Good	J	336

[11]:
```
# Select another set of variables from the df table

df2 = df %>% select(depth,table,x,z)
head(df2)
```

depth	table	x	z
61.5	55	3.95	2.43
59.8	61	3.89	2.31
56.9	65	4.05	2.31
62.4	58	4.20	2.63
63.3	58	4.34	2.75
62.8	57	3.94	2.48

0.0.4 Using the Summarize function

[12]:
```
summarize(df,pc=mean(price))
```

pc
3932.8

[13]:
```
# Summarize the table column of df2

summarize(df2, TB=mean(table))
```

TB
57.45718

0.0.5 Using the Groupby function

[14]:
```
df %>%
group_by(cut) %>%
summarize(average_price = mean(price))
```

cut	average_price
Fair	4358.758
Good	3928.864
Very Good	3981.760
Premium	4584.258
Ideal	3457.542

0.0.6 Using arrange to order data

[15]:
```
# given our data frame df

head(df)
```

carat	cut	color	clarity	depth	table	price	x	y	z
0.23	Ideal	E	SI2	61.5	55	326	3.95	3.98	2.43
0.21	Premium	E	SI1	59.8	61	326	3.89	3.84	2.31
0.23	Good	E	VS1	56.9	65	327	4.05	4.07	2.31
0.29	Premium	I	VS2	62.4	58	334	4.20	4.23	2.63
0.31	Good	J	SI2	63.3	58	335	4.34	4.35	2.75
0.24	Very Good	J	VVS2	62.8	57	336	3.94	3.96	2.48

[20]:
```
# we wan to reorder the df in the order of price, carat, table

```

```
rangedf <- df %>% arrange(price,carat)
head(rangedf)
```

carat	cut	color	clarity	depth	table	price	x	y	z
0.21	Premium	E	SI1	59.8	61	326	3.89	3.84	2.31
0.23	Ideal	E	SI2	61.5	55	326	3.95	3.98	2.43
0.23	Good	E	VS1	56.9	65	327	4.05	4.07	2.31
0.29	Premium	I	VS2	62.4	58	334	4.20	4.23	2.63
0.31	Good	J	SI2	63.3	58	335	4.34	4.35	2.75
0.24	Very Good	J	VVS2	62.8	57	336	3.94	3.96	2.48

0.0.7 Using mutate to add new variables

[21]:
```
# Add a total revenue column by multiplying price by depth

Total_Revenu = df %>% mutate(TR = price*depth)
head(Total_Revenu)
```

carat	cut	color	clarity	depth	table	price	x	y	z	TR
0.23	Ideal	E	SI2	61.5	55	326	3.95	3.98	2.43	20049.0
0.21	Premium	E	SI1	59.8	61	326	3.89	3.84	2.31	19494.8
0.23	Good	E	VS1	56.9	65	327	4.05	4.07	2.31	18606.3
0.29	Premium	I	VS2	62.4	58	334	4.20	4.23	2.63	20841.6
0.31	Good	J	SI2	63.3	58	335	4.34	4.35	2.75	21205.5
0.24	Very Good	J	VVS2	62.8	57	336	3.94	3.96	2.48	21100.8

0.0.8 Using RBIND and CBIND functions

[22]:
```
# Splitting the df dataset into 2 df_1 and df_2
df_1 <- df %>% select(carat,cut,color,clarity,depth,table)
df_2 <- df %>% select(price,x,y,z)
```

[25]:
```
head(df_1)
```

carat	cut	color	clarity	depth	table
0.23	Ideal	E	SI2	61.5	55
0.21	Premium	E	SI1	59.8	61
0.23	Good	E	VS1	56.9	65
0.29	Premium	I	VS2	62.4	58
0.31	Good	J	SI2	63.3	58
0.24	Very Good	J	VVS2	62.8	57

[26]:
```
head(df_2)
```

price	x	y	z
326	3.95	3.98	2.43
326	3.89	3.84	2.31
327	4.05	4.07	2.31
334	4.20	4.23	2.63
335	4.34	4.35	2.75
336	3.94	3.96	2.48

[27]:
```
# cbind df_2 to df_1

new_df <- cbind(df_1,df_2)
head(new_df)
```

carat	cut	color	clarity	depth	table	price	x	y	z
0.23	Ideal	E	SI2	61.5	55	326	3.95	3.98	2.43
0.21	Premium	E	SI1	59.8	61	326	3.89	3.84	2.31
0.23	Good	E	VS1	56.9	65	327	4.05	4.07	2.31
0.29	Premium	I	VS2	62.4	58	334	4.20	4.23	2.63
0.31	Good	J	SI2	63.3	58	335	4.34	4.35	2.75
0.24	Very Good	J	VVS2	62.8	57	336	3.94	3.96	2.48

[28]:
```
# filter some rows and rbind it to new_df

prem_df <- filter(df,cut=="Premium")
head(prem_df)
```

carat	cut	color	clarity	depth	table	price	x	y	z
0.21	Premium	E	SI1	59.8	61	326	3.89	3.84	2.31
0.29	Premium	I	VS2	62.4	58	334	4.20	4.23	2.63
0.22	Premium	F	SI1	60.4	61	342	3.88	3.84	2.33
0.20	Premium	E	SI2	60.2	62	345	3.79	3.75	2.27
0.32	Premium	E	I1	60.9	58	345	4.38	4.42	2.68
0.24	Premium	I	VS1	62.5	57	355	3.97	3.94	2.47

[29]:
```
new_df2 <- rbind(new_df,prem_df)
```

[30]:
```
str(new_df)
str(new_df2)
```

```
'data.frame':   53940 obs. of  10 variables:
 $ carat  : num  0.23 0.21 0.23 0.29 0.31 0.24 0.24 0.26 0.22 0.23 ...
 $ cut    : Ord.factor w/ 5 levels "Fair"<"Good"<..: 5 4 2 4 2 3 3 3 1 3 ...
 $ color  : Ord.factor w/ 7 levels "D"<"E"<"F"<"G"<..: 2 2 2 6 7 7 6 5 2 5 ...
 $ clarity: Ord.factor w/ 8 levels "I1"<"SI2"<"SI1"<..: 2 3 5 4 2 6 7 3 4 5 ...
 $ depth  : num  61.5 59.8 56.9 62.4 63.3 62.8 62.3 61.9 65.1 59.4 ...
 $ table  : num  55 61 65 58 58 57 57 55 61 61 ...
 $ price  : int  326 326 327 334 335 336 336 337 337 338 ...
 $ x      : num  3.95 3.89 4.05 4.2 4.34 3.94 3.95 4.07 3.87 4 ...
 $ y      : num  3.98 3.84 4.07 4.23 4.35 3.96 3.98 4.11 3.78 4.05 ...
 $ z      : num  2.43 2.31 2.31 2.63 2.75 2.48 2.47 2.53 2.49 2.39 ...
'data.frame':   67731 obs. of  10 variables:
 $ carat  : num  0.23 0.21 0.23 0.29 0.31 0.24 0.24 0.26 0.22 0.23 ...
 $ cut    : Ord.factor w/ 5 levels "Fair"<"Good"<..: 5 4 2 4 2 3 3 3 1 3 ...
 $ color  : Ord.factor w/ 7 levels "D"<"E"<"F"<"G"<..: 2 2 2 6 7 7 6 5 2 5 ...
 $ clarity: Ord.factor w/ 8 levels "I1"<"SI2"<"SI1"<..: 2 3 5 4 2 6 7 3 4 5 ...
 $ depth  : num  61.5 59.8 56.9 62.4 63.3 62.8 62.3 61.9 65.1 59.4 ...
 $ table  : num  55 61 65 58 58 57 57 55 61 61 ...
 $ price  : int  326 326 327 334 335 336 336 337 337 338 ...
 $ x      : num  3.95 3.89 4.05 4.2 4.34 3.94 3.95 4.07 3.87 4 ...
```

```
$ y       : num  3.98 3.84 4.07 4.23 4.35 3.96 3.98 4.11 3.78 4.05 ...
$ z       : num  2.43 2.31 2.31 2.63 2.75 2.48 2.47 2.53 2.49 2.39 ...
```

0.0.9 String Operations

[31]:
```
require(stringr)
```

```
Loading required package: stringr
Warning message:
```

[32]:
```
x <- 'Notice that spaces were put between the strings'
```

[33]:
```
# split x

splt <- strsplit(x," ")
print(splt)
```

```
[[1]]
[1] "Notice"  "that"    "spaces"  "were"    "put"     "between" "the"
[8] "strings"
```

[36]:
```
# Concatenating strings

a <- 'competitive'
b <- 'coding'
c <- 'is difficult'

abc <- cat(a,b,c,sep =' ')
```

[ ]:
```

```