PartC_Mid Sem 1 Solution

phizer Datset -Payments made by Pfizer to doctors across the United States in the second half on 2009.[All question carry 5 marks each]

```
library(tidyverse)
```

Question 1 . Answer the following with respect to the dataset shown below. Write the corresponding R code for all the questions.

```
## Warning in system("timedatectl", intern = TRUE): running command 'timedatectl'
## had status 1
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.2 --v ggplot2 3.3.6
## v tibble  3.1.8      v dplyr   1.0.10
## v tidyr   1.2.0      v stringr 1.4.0
## v readr   2.1.2      v forcats 0.5.1 -- Conflicts ------------------------------------------ tidyvers
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(readr)
phzr <- read.csv("pfizer.csv")
```

```
head(phzr,10)
```

```
##                     org_indiv       first_plus first_name   last_name
## 1    3-D MEDICAL SERVICES LLC     STEVEN BRUCE     STEVEN DEITELZWEIG
## 2           AA DOCTORS, INC.     AAKASH MOHAN     AAKASH       AHUJA
## 3      ABBO, LILIAN MARGARITA LILIAN MARGARITA     LILIAN        ABBO
## 4      ABBO, LILIAN MARGARITA LILIAN MARGARITA     LILIAN        ABBO
## 5      ABBO, LILIAN MARGARITA LILIAN MARGARITA     LILIAN        ABBO
## 6       ABDULLAH RAFFEE MD PC         ABDULLAH   ABDULLAH      RAFFEE
## 7             ABEBE, SHEILA Y         SHEILA Y     SHEILA       ABEBE
## 8             ABEBE, SHEILA Y         SHEILA Y     SHEILA       ABEBE
## 9  ABILENE FAMILY FOOT CENTER      GALEN CHRIS      GALEN   ALBRITTON
## 10            ABOLNIK, IGOR Z           IGOR Z       IGOR     ABOLNIK
##            city state                category cash other total
## 1   NEW ORLEANS    LA   Professional Advising 2625     0  2625
## 2   PASO ROBLES    CA       Expert-Led Forums 1000     0  1000
## 3         MIAMI    FL Business Related Travel    0   448   448
## 4         MIAMI    FL                   Meals    0   119   119
## 5         MIAMI    FL   Professional Advising 1800     0  1800
## 6         FLINT    MI       Expert-Led Forums  750     0   750
## 7  INDIANAPOLIS    IN       Educational Items    0    47    47
## 8  INDIANAPOLIS    IN       Expert-Led Forums  825     0   825
## 9       ABILENE    TX   Professional Advising 3000     0  3000
## 10        PROVO    UT Business Related Travel    0   396   396
```

```
str(phzr)
```

```
## 'data.frame':    10087 obs. of  10 variables:
##  $ org_indiv : chr  "3-D MEDICAL SERVICES LLC" "AA DOCTORS, INC." "ABBO, LILIAN MARGARITA" "ABBO, LI
##  $ first_plus: chr  "STEVEN BRUCE" "AAKASH MOHAN" "LILIAN MARGARITA" "LILIAN MARGARITA" ...
##  $ first_name: chr  "STEVEN" "AAKASH" "LILIAN" "LILIAN" ...
##  $ last_name : chr  "DEITELZWEIG" "AHUJA" "ABBO" "ABBO" ...
##  $ city      : chr  "NEW ORLEANS" "PASO ROBLES" "MIAMI" "MIAMI" ...
##  $ state     : chr  "LA" "CA" "FL" "FL" ...
##  $ category  : chr  "Professional Advising" "Expert-Led Forums" "Business Related Travel" "Meals" ..
##  $ cash      : int  2625 1000 0 0 1800 750 0 825 3000 0 ...
##  $ other     : int  0 0 448 119 0 0 47 0 0 396 ...
##  $ total     : int  2625 1000 448 119 1800 750 47 825 3000 396 ...
```

1.Find doctors in MIAMI paid $3500 or more by Pfizer to run category Professional Advising.

```
expert_1000 <- phzr %>%
  filter(city == "MIAMI" & total >= 3500 & category == "Professional Advising")
expert_1000
```

2.Sort the list of doctors in city "INDIANAPOLIS",who run for category "Expert-Led Forums" in descending order by the payments received .

```
sort_doc <- phzr %>%
  filter(city=="INDIANAPOLIS" & category == "Expert-Led Forums") %>%
  arrange(desc(total))

sort_doc
```

3.Find doctors in California (CA) or Florida (FL) who were paid $10,000 or more by Pfizer.

```
ca_fl_expert_10000 <- phzr %>%
  filter((state == "CA" | state == "FL") & total >= 10000)

ca_fl_expert_10000
```

4.For each state,Calculate the total payments, median payments and the number of payments.

```
state_summary <- phzr %>%
  group_by(state) %>%
  summarize(sum = sum(total), median = median(total), count = n())
state_summary
```

5.Filter the data for all payments for categories Expert-Led Forums or Business Related Travel, and arrange alphabetically by doctor's last name first, then first name.

```
expert_advice <- phzr %>%
  filter(category == "Expert-Led Forums" | category == "Business Related Travel") %>%
  arrange(last_name, first_name)
```

Cereal Dataset-This dataset contains nutrition information for 77 breakfast cereals and includes 16 variables.

```
cereal <- read.csv("cereal.csv")
```

```
head(cereal,10)
```

Question 2. Answer the following with respect to the dataset shown below. Write the corresponding R code for all the questions.

```
##                         name mfr type calories protein fat sodium fiber carbo
## 1                  100% Bran   N    C       70       4   1    130  10.0   5.0
## 2          100% Natural Bran   Q    C      120       3   5     15   2.0   8.0
## 3                   All-Bran   K    C       70       4   1    260   9.0   7.0
## 4  All-Bran with Extra Fiber   K    C       50       4   0    140  14.0   8.0
## 5             Almond Delight   R    C      110       2   2    200   1.0  14.0
## 6    Apple Cinnamon Cheerios   G    C      110       2   2    180   1.5  10.5
## 7                 Apple Jacks   K    C      110       2   0    125   1.0  11.0
## 8                     Basic 4   G    C      130       3   2    210   2.0  18.0
## 9                   Bran Chex   R    C       90       2   1    200   4.0  15.0
## 10                Bran Flakes   P    C       90       3   0    210   5.0  13.0
##    sugars potass vitamins shelf weight cups   rating
## 1       6    280       25     3   1.00 0.33 68.40297
## 2       8    135        0     3   1.00 1.00 33.98368
## 3       5    320       25     3   1.00 0.33 59.42551
## 4       0    330       25     3   1.00 0.50 93.70491
## 5       8     -1       25     3   1.00 0.75 34.38484
## 6      10     70       25     1   1.00 0.75 29.50954
## 7      14     30       25     2   1.00 1.00 33.17409
## 8       8    100       25     3   1.33 0.75 37.03856
## 9       6    125       25     1   1.00 0.67 49.12025
## 10      5    190       25     3   1.00 0.67 53.31381
```

```
str(cereal)
```

```
## 'data.frame':    77 obs. of  16 variables:
##  $ name    : chr  "100% Bran" "100% Natural Bran" "All-Bran" "All-Bran with Extra Fiber" ...
##  $ mfr     : chr  "N" "Q" "K" "K" ...
##  $ type    : chr  "C" "C" "C" "C" ...
##  $ calories: int  70 120 70 50 110 110 110 130 90 90 ...
##  $ protein : int  4 3 4 4 2 2 2 3 2 3 ...
##  $ fat     : int  1 5 1 0 2 2 0 2 1 0 ...
##  $ sodium  : int  130 15 260 140 200 180 125 210 200 210 ...
##  $ fiber   : num  10 2 9 14 1 1.5 1 2 4 5 ...
##  $ carbo   : num  5 8 7 8 14 10.5 11 18 15 13 ...
##  $ sugars  : int  6 8 5 0 8 10 14 8 6 5 ...
##  $ potass  : int  280 135 320 330 -1 70 30 100 125 190 ...
##  $ vitamins: int  25 0 25 25 25 25 25 25 25 25 ...
##  $ shelf   : int  3 3 3 3 3 1 2 3 1 3 ...
##  $ weight  : num  1 1 1 1 1 1 1 1.33 1 1 ...
##  $ cups    : num  0.33 1 0.33 0.5 0.75 0.75 1 0.75 0.67 0.67 ...
##  $ rating  : num  68.4 34 59.4 93.7 34.4 ...
```

1 Add a new variable to the dataset called 'totalcarbo', which is the sum of 'carbo' and 'sugars'. (*2 marks*)

```
totalcarbo <- mutate(cereal, totalcarbo = carbo + sugars)
cereal
```

2.Find out the count of cold(C) cereal and sort it in descending order according to calories. (*3 marks*)

```
  cold_cereals <- filter(cereal, type=="C")
  nrow(cold_cereals)
  cereal %>% arrange(desc(calories))
```

3.Create a subset of the dataframe containing cereals that contain at least 1 unit of sugar, and keep only the variables 'name', 'calories' and 'carbo'. Then inspect the first few rows of the dataframe.(*2 marks*)

```
# solution 1

select(filter(cereal,sugars >= 1),name, calories, carbo)
```

4.Get a subset of the dataframe of all cereals having 'carbo' between 8 and 12 units inclusive.(*2 marks*)

```
cereal %>%
filter(carbo >= 8 & carbo <=12)
```

5.Determine the number of distinct cereal Manufactures in the data set and calculate the total mean average of variables sodium, fiber,sugar,vitamins and carbohydrate and store it in a variable 'avg_length'. (*5 marks*)

```
length(unique(cereal$mfr))
cereal %>%
summarise(avg_length=mean(c(sodium,fiber,sugars,vitamins,carbo))) ->cereal_avg
cereal_avg
```

6. Add a new variable 'carbo_class', which is 'high' when totalcarbo >15 and 'low' otherwise. Make sure the new variable is a factor. Also,find the average, minimum and maximum sugar content for 'low' and 'high' carbo. (*6 marks*)

```
#solution 1
cereal$carbo_Class <- factor(ifelse(cereal$totalcarbo > 15,"high","low"))
cereal
with(cereal, tapply(sugars,carbo_Class,min,na.rm=TRUE))

with(cereal, tapply(sugars,carbo_Class,max,na.rm=TRUE))

with(cereal, tapply(sugars,carbo_Class,mean,na.rm=TRUE))
```

```
#solution 2

carbo_class <- mutate(cereal,ifelse(cereal$totalcarbo > 15,"high","low"))
summarise(cereal, mean(sugars),min(sugars),max(sugars))
cereal
```