IT 497 Exam 2 Question 2

# Question 2.   
library(tidyverse)

## -- Attaching packages -------------------------------------------- tidyverse 1.3.0 --

## v ggplot2 3.3.2 v purrr 0.3.4  
## v tibble 3.0.3 v dplyr 1.0.2  
## v tidyr 1.1.2 v stringr 1.4.0  
## v readr 1.3.1 v forcats 0.5.0

## -- Conflicts ----------------------------------------------- tidyverse\_conflicts() --  
## x dplyr::filter() masks stats::filter()  
## x dplyr::lag() masks stats::lag()

library(readr)  
  
# a. Read in the following Milk Cow data:  
df <- read\_csv("https://github.com/rfordatascience/tidytuesday/raw/master/data/2019/2019-01-29/milkcow\_facts.csv")

## Parsed with column specification:  
## cols(  
## year = col\_double(),  
## avg\_milk\_cow\_number = col\_double(),  
## milk\_per\_cow = col\_double(),  
## milk\_production\_lbs = col\_double(),  
## avg\_price\_milk = col\_double(),  
## dairy\_ration = col\_double(),  
## milk\_feed\_price\_ratio = col\_double(),  
## milk\_cow\_cost\_per\_animal = col\_double(),  
## milk\_volume\_to\_buy\_cow\_in\_lbs = col\_double(),  
## alfalfa\_hay\_price = col\_double(),  
## slaughter\_cow\_price = col\_double()  
## )

# b. Examine the data (look at head and tail)  
# Head  
head(df)

## # A tibble: 6 x 11  
## year avg\_milk\_cow\_nu~ milk\_per\_cow milk\_production~ avg\_price\_milk  
## <dbl> <dbl> <dbl> <dbl> <dbl>  
## 1 1980 10799000 11891 128406000000 0.13   
## 2 1981 10898000 12183 132770000000 0.138  
## 3 1982 11011000 12306 135505000000 0.136  
## 4 1983 11059000 12622 139588000000 0.136  
## 5 1984 10793000 12541 135351000000 0.135  
## 6 1985 10981000 13024 143012000000 0.127  
## # ... with 6 more variables: dairy\_ration <dbl>, milk\_feed\_price\_ratio <dbl>,  
## # milk\_cow\_cost\_per\_animal <dbl>, milk\_volume\_to\_buy\_cow\_in\_lbs <dbl>,  
## # alfalfa\_hay\_price <dbl>, slaughter\_cow\_price <dbl>

# Tail  
tail(df)

## # A tibble: 6 x 11  
## year avg\_milk\_cow\_nu~ milk\_per\_cow milk\_production~ avg\_price\_milk  
## <dbl> <dbl> <dbl> <dbl> <dbl>  
## 1 2009 9202000 20561 189202000000 0.128  
## 2 2010 9123000 21142 192877000000 0.163  
## 3 2011 9199000 21334 196255000000 0.201  
## 4 2012 9237000 21722 200642000000 0.185  
## 5 2013 9224000 21816 201231000000 0.201  
## 6 2014 9257000 22259 206054000000 0.24   
## # ... with 6 more variables: dairy\_ration <dbl>, milk\_feed\_price\_ratio <dbl>,  
## # milk\_cow\_cost\_per\_animal <dbl>, milk\_volume\_to\_buy\_cow\_in\_lbs <dbl>,  
## # alfalfa\_hay\_price <dbl>, slaughter\_cow\_price <dbl>

# c. Examine the structure (str) of the data  
str(df)

## tibble [35 x 11] (S3: spec\_tbl\_df/tbl\_df/tbl/data.frame)  
## $ year : num [1:35] 1980 1981 1982 1983 1984 ...  
## $ avg\_milk\_cow\_number : num [1:35] 10799000 10898000 11011000 11059000 10793000 ...  
## $ milk\_per\_cow : num [1:35] 11891 12183 12306 12622 12541 ...  
## $ milk\_production\_lbs : num [1:35] 1.28e+11 1.33e+11 1.36e+11 1.40e+11 1.35e+11 ...  
## $ avg\_price\_milk : num [1:35] 0.13 0.138 0.136 0.136 0.135 0.127 0.125 0.125 0.122 0.136 ...  
## $ dairy\_ration : num [1:35] 0.0484 0.0504 0.0442 0.0524 0.0534 ...  
## $ milk\_feed\_price\_ratio : num [1:35] 2.72 2.76 3.09 2.61 2.54 ...  
## $ milk\_cow\_cost\_per\_animal : num [1:35] 1190 1200 1110 1030 895 860 820 920 990 1030 ...  
## $ milk\_volume\_to\_buy\_cow\_in\_lbs: num [1:35] 9154 8696 8162 7574 6630 ...  
## $ alfalfa\_hay\_price : num [1:35] 72 70.9 72.7 78.7 79.5 ...  
## $ slaughter\_cow\_price : num [1:35] 0.457 0.419 0.4 0.394 0.398 ...  
## - attr(\*, "spec")=  
## .. cols(  
## .. year = col\_double(),  
## .. avg\_milk\_cow\_number = col\_double(),  
## .. milk\_per\_cow = col\_double(),  
## .. milk\_production\_lbs = col\_double(),  
## .. avg\_price\_milk = col\_double(),  
## .. dairy\_ration = col\_double(),  
## .. milk\_feed\_price\_ratio = col\_double(),  
## .. milk\_cow\_cost\_per\_animal = col\_double(),  
## .. milk\_volume\_to\_buy\_cow\_in\_lbs = col\_double(),  
## .. alfalfa\_hay\_price = col\_double(),  
## .. slaughter\_cow\_price = col\_double()  
## .. )

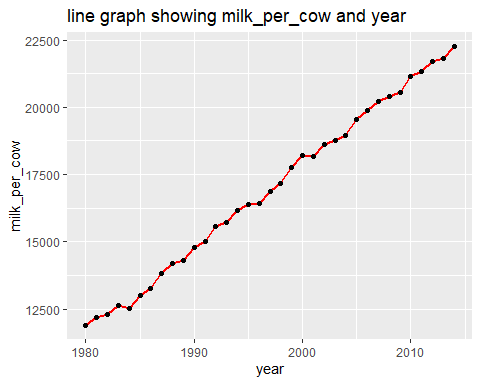
# d. Use select to obtain the columns: year, avg\_milk\_cow\_number, milk\_per\_cow, avg\_price\_milk, milk\_cow\_cost\_per\_animal, alfalfa\_hay\_price and slaughter\_cow\_price.  
  
df1 <- select(df, year, avg\_milk\_cow\_number, milk\_per\_cow, avg\_price\_milk, milk\_cow\_cost\_per\_animal, alfalfa\_hay\_price, slaughter\_cow\_price)  
df1

## # A tibble: 35 x 7  
## year avg\_milk\_cow\_nu~ milk\_per\_cow avg\_price\_milk milk\_cow\_cost\_p~  
## <dbl> <dbl> <dbl> <dbl> <dbl>  
## 1 1980 10799000 11891 0.13 1190  
## 2 1981 10898000 12183 0.138 1200  
## 3 1982 11011000 12306 0.136 1110  
## 4 1983 11059000 12622 0.136 1030  
## 5 1984 10793000 12541 0.135 895  
## 6 1985 10981000 13024 0.127 860  
## 7 1986 10773000 13285 0.125 820  
## 8 1987 10327000 13819 0.125 920  
## 9 1988 10224000 14185 0.122 990  
## 10 1989 10046000 14323 0.136 1030  
## # ... with 25 more rows, and 2 more variables: alfalfa\_hay\_price <dbl>,  
## # slaughter\_cow\_price <dbl>

# e. Omit all other columns  
df <- select(df, -c(milk\_production\_lbs, dairy\_ration, milk\_feed\_price\_ratio, milk\_volume\_to\_buy\_cow\_in\_lbs))  
df

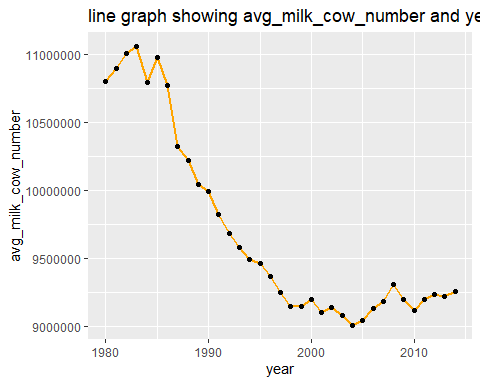
## # A tibble: 35 x 7  
## year avg\_milk\_cow\_nu~ milk\_per\_cow avg\_price\_milk milk\_cow\_cost\_p~  
## <dbl> <dbl> <dbl> <dbl> <dbl>  
## 1 1980 10799000 11891 0.13 1190  
## 2 1981 10898000 12183 0.138 1200  
## 3 1982 11011000 12306 0.136 1110  
## 4 1983 11059000 12622 0.136 1030  
## 5 1984 10793000 12541 0.135 895  
## 6 1985 10981000 13024 0.127 860  
## 7 1986 10773000 13285 0.125 820  
## 8 1987 10327000 13819 0.125 920  
## 9 1988 10224000 14185 0.122 990  
## 10 1989 10046000 14323 0.136 1030  
## # ... with 25 more rows, and 2 more variables: alfalfa\_hay\_price <dbl>,  
## # slaughter\_cow\_price <dbl>

# f1. Using a line graph, show milk\_per\_cow and year.  
library(ggplot2)  
  
line\_plot = ggplot(data = df, aes(x = year, y = milk\_per\_cow)) + geom\_line(color= "red", size=1) + geom\_point() + ggtitle("line graph showing milk\_per\_cow and year")  
  
line\_plot



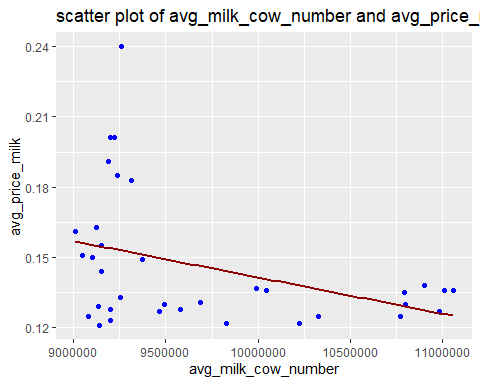
# f2. Write a two sentence explanation of the graph.  
  
# This is a line graph showing the year of milk production and the amount of milk per cow for the given years, where the years are on the abcissor and the milk per cow on the y-axis.  
  
# In this graph, we can see that as the years increase from 1980 to 2010, the amount of milk produced per cow seem to increase with small amount of decrement in some years. Example, there was a smalle decrease from 1983 to 1984, 1988 to 1989, 1995 to 1996, etc. This graph shows a positive relationship betweeen year and milk per cow.

# g1. Using a line graph, show avg\_milk\_cow\_number and year.  
line\_plot2 = ggplot(data = df, aes(x = year, y = avg\_milk\_cow\_number)) + geom\_line(color= "orange", size=1) + geom\_point() + ggtitle("line graph showing avg\_milk\_cow\_number and year")  
  
line\_plot2



# g2. Write a two sentence explanation of the graph above.  
  
# This graph shows the years from 1980 to 2010 of cow milk production and the average milk cow number, where the years are on the abscessor and avgerage milk cow number on the y-axis.   
  
# From this graph, the average milk cow number increases initially from 1980 with about 10800000 average milk cow numbers, reaching the maximum at 1983 with about 11050000 and then begun to fall to 1984 with about 10800000 and rise again from 1984 to 1985 with nearly 11000000 average milk cow number. From 1985, it fell reaching the minimum at 2004 with 9000000 and then rises again to 2008 with about 9300000. It then fell from 2008 to 2010 and then began to rise again. We may say there is an irregular flow between the two variables.

# h1. Find the scatter plot of avg\_milk\_cow\_number and avg\_price\_milk.  
ggplot(df, aes(x = avg\_milk\_cow\_number, y = avg\_price\_milk)) + geom\_point(color="blue") + geom\_smooth(method = lm, formula = y ~ x, se = FALSE, color="darkred", fill="blue") + ggtitle("scatter plot of avg\_milk\_cow\_number and avg\_price\_milk")



# h2. Does the plot reveal any relationship between the variables? Write a two sentence explanation of the graph.  
  
# No. This graph does not reveal a relationship between the variables.  
  
# From this graph, we can see that the average milk cow number on the x-axis is plotted against the average price milk with the data points scatterd without a particular patten.   
  
# We can also see that most data points do not fall close to the line of best fit showing a considereably high vriablility between the data points and the line. These show no patten of relationship existing between the two variables.