STAT 3675Q Homework 3

Due date: **Thursday, September 18, at noon**

Zeshi Feng

Note:

- Ensure that your code is fully visible in the PDF and not cropped. If needed, break the code into multiple lines to fit.
- It is recommended to write descriptive answers outside of R code chunks (i.e., as text in the main body), while comments within the code chunks can be reserved for brief code annotations.
- In all homework questions, include a written explanation of any output to earn full credit.

Question 1 [20 points]

a. Create vectors with the following names and elements.

- Subject: Math, Science, History, Music
- Midterm: $95, 87, 39, 67$
- Final: $93, 90, 32, 88$
- Grade: A, B, c, B

Answer:

```
Subject <- c("Math", "Science", "History", "Music")
Midterm <- c(95, 87, 39, 67)
Final   <- c(93, 90, 32, 88)
Grade   <- c("A", "B", "C", "B")

Subject
```

```
## [1] "Math"    "Science" "History" "Music"
```

```
Midterm
```

```
## [1] 95 87 39 67
```

```
Final
```

```
## [1] 93 90 32 88
```

```
Grade
```

```
## [1] "A" "B" "C" "B"
```

b. Convert Grade to an ordered factor (ordinal variable) with levels C<B<A. Then create a data frame containing the four variables created in part a.

Answer:

```
Grade <- factor(c("A", "B", "C", "B"),
                levels = c("C", "B", "A"),
                ordered = TRUE)
results <- data.frame(Subject, Midterm, Final, Grade)
results
```

```
##   Subject Midterm Final Grade
## 1    Math      95    93     A
## 2 Science      87    90     B
## 3 History      39    32     C
## 4   Music      67    88     B
```

c. Suppose that the data above are grades of a student named Katty in 2022. Create a list for Katty's grade as follows.

```
## $name
## [1] "Katty"
##
## $year
## [1] 2022
##
## $score
## Subject Midterm Final Grade
## 1 Math 95 93 A
## 2 Science 87 90 B
## 3 History 39 32 C
## 4 Music 67 88 B
```

Answer:

```
Katty_grade <- list(
  name = "Katty",
  year = 2022,
  score = results
)
Katty_grade
```

```
## $name
## [1] "Katty"
```

```
##
## $year
## [1] 2022
##
## $score
##   Subject Midterm Final Grade
## 1    Math      95    93     A
## 2 Science      87    90     B
## 3 History      39    32     C
## 4   Music      67    88     B
```

d. What is Katty's History grade? Use the function `which()`.

Answer:

```
row_index <- which(Katty_grade$score$Subject == "History")
Katty_grade$score$Grade[row_index]
```

```
## [1] C
## Levels: C < B < A
```

"grade is C"

Question 2 [30 points]

a. Create the following data frame:

Cereal.name	Manufacturer	Cold.or.Hot	calories	rating
100%_Bran	N	C	70	68.4
100%_Natural_Bran	Q	C	120	34.0
All-Bran	K	H	70	59.4
All-Bran_with_Extra_Fiber	K	C	50	93.7
Almond_Delight	R	H	110	34.4
Apple_Cinnamon_Cheerios	G	C	110	29.5

where `Manufacturer` and `Cold.or.Hot` should be created as factors, and `Cereal.name` should be used as the case identifier.

```
# The following is given for your convenience
Cereal.name <- c("100%_Bran", "100%_Natural_Bran", "All-Bran",
                 "All-Bran_with_Extra_Fiber", "Almond_Delight",
                 "Apple_Cinnamon_Cheerios")
Manufacturer <- c("N", "Q", "K", "K", "R", "G")
Cold.or.Hot <- c("C", "C", "H", "C", "H", "C")
calories <- c(70, 120, 70, 50, 110, 110)
rating <- c(68.4, 34, 59.4, 93.7, 34.4, 29.5)
```

Answer:

```
cereal_df <- data.frame(Manufacturer, Cold.or.Hot, calories, rating, row.names = Cereal.
cereal_df
```

```
##                            Manufacturer Cold.or.Hot calories rating
## 100%_Bran                              N           C       70   68.4
## 100%_Natural_Bran                      Q           C      120   34.0
## All-Bran                               K           H       70   59.4
## All-Bran_with_Extra_Fiber              K           C       50   93.7
## Almond_Delight                         R           H      110   34.4
## Apple_Cinnamon_Cheerios                G           C      110   29.5
```

b. Create an ordered factor named **grade** in the data frame, which takes the value "high" if **rating** is greater than 90, "low" if **rating** is less than 40, and "median" otherwise. The order of the levels should be low=1, median=2, high=3.

Answer:

```
grade <- ifelse(cereal_df$rating > 90, "high",
                ifelse(cereal_df$rating < 40, "low", "median"))
grade <- factor(grade,
                levels = c("low", "median", "high"),
                ordered = TRUE)
cereal_df$grade <- grade
cereal_df
```

```
##                            Manufacturer Cold.or.Hot calories rating  grade
## 100%_Bran                              N           C       70   68.4 median
## 100%_Natural_Bran                      Q           C      120   34.0    low
## All-Bran                               K           H       70   59.4 median
## All-Bran_with_Extra_Fiber              K           C       50   93.7   high
## Almond_Delight                         R           H      110   34.4    low
## Apple_Cinnamon_Cheerios                G           C      110   29.5    low
```

c. Extract the manufacturer and calories information of the all cereals with low grade.

Answer:

```
low_grade_info <- cereal_df[cereal_df$grade == "low",
                            c("Manufacturer", "calories")]
low_grade_info
```

```
##                          Manufacturer calories
## 100%_Natural_Bran                   Q      120
## Almond_Delight                      R      110
## Apple_Cinnamon_Cheerios             G      110
```

d. Create a table that displays the count of occurrences for each combination of manufacturer and grade.

Answer:

```
table_manuf_grade <- table(cereal_df$Manufacturer, cereal_df$grade)
print(table_manuf_grade)
```

```
## 
##     low median high
##   G   1      0    0
##   K   0      1    1
##   N   0      1    0
##   Q   1      0    0
##   R   1      0    0
```

e. Create a list containing information about cold cereals. Include the following three components in the list: `Cereal.name`, `Manufacturer`, and `rating`.

Answer:

```
cold_cereals <- cereal_df[cereal_df$Cold.or.Hot == "C", ]
cold_list <- list(
  Cereal.name = rownames(cold_cereals),
  Manufacturer = cold_cereals$Manufacturer,
  rating = cold_cereals$rating
)
cold_list
```

```
## $Cereal.name
## [1] "100%_Bran"                 "100%_Natural_Bran"
## [3] "All-Bran_with_Extra_Fiber" "Apple_Cinnamon_Cheerios"
## 
## $Manufacturer
## [1] "N" "Q" "K" "G"
## 
## $rating
## [1] 68.4 34.0 93.7 29.5
```

Question 3 [50 points]

Data: The Forbes Global 2000 list is a ranking of the world's biggest companies, measured by sales, profits, assets and market value (Year 2014). (http://www.forbes.com/global2000/list/#tab:overall)

Here is a description of the columns in the data set

- Company: the name of the company
- Sector: a factor describing the products the company produces

- Industry: a factor giving the industry the company belongs to
- Continent: a factor giving the continent the company is situated in
- Country: a factor giving the country the company is situated in
- Market Value: the market value of the company in billion USD
- Sales: the amount of sales of the company in billion USD
- Profits: the profit of the company in billion USD
- Assets: the assets of the company in billion USD
- Rank: the ranking of the company
- Forbes Webpage: a character string describing webpage whitin Forbes.com

a. Use `read.csv()` to load the data set `Forbes Global 2000.csv` into R and store it as **ForbesGlobal2000**.

Answer:

```
ForbesGlobal2000 <- read.csv("Forbes Global 2000.csv")
```

b. Display the first and last few records in the data using the functions `head()` and `tail()`, respectively.

Answer:

```
head(ForbesGlobal2000)
```

```
##                       Company     Sector             Industry     Continent
## 1                       ICBC Financials          Major Banks          Asia
## 2    China Construction Bank Financials       Regional Banks          Asia
## 3 Agricultural Bank of China Financials       Regional Banks          Asia
## 4             JPMorgan Chase Financials          Major Banks North America
## 5         Berkshire Hathaway Financials  Investment Services North America
## 6                Exxon Mobil     Energy Oil & Gas Operations North America
##         Country Market.Value Sales Profits Assets Rank
## 1         China        215.6 148.7    42.7 3124.9    1
## 2         China        174.4 121.3    34.2 2449.5    2
## 3         China        141.1 136.4    27.0 2405.4    3
## 4 United States        229.7 105.7    17.3 2435.3    4
## 5 United States        309.1 178.8    19.5  493.4    5
## 6 United States        422.3 394.0    32.6  346.8    6
##                                                Forbes.Webpage
## 1                       http://www.forbes.com/companies/icbc/
## 2    http://www.forbes.com/companies/china-construction-bank/
## 3 http://www.forbes.com/companies/agricultural-bank-of-china/
## 4             http://www.forbes.com/companies/jpmorgan-chase/
## 5         http://www.forbes.com/companies/berkshire-hathaway/
## 6                http://www.forbes.com/companies/exxon-mobil/
```

```
tail(ForbesGlobal2000)
```

```
##                 Company                 Sector                     Industry
## 1995       Shikoku Bank             Financials               Regional Banks
## 1996             Cameco              Materials  Diversified Metals & Mining
## 1997          BMCE Bank             Financials               Regional Banks
## 1998  Synovus Financial             Financials               Regional Banks
## 1999            Equifax Consumer Discretionary Business & Personal Services
## 2000 UNY Group Holdings Consumer Discretionary             Specialty Stores
##          Continent       Country Market.Value Sales Profits Assets Rank
## 1995           Asia         Japan          0.4   0.4     0.1   26.5 1994
## 1996 North America        Canada          9.5   2.4     0.3    7.6 1996
## 1997        Africa       Morocco          4.6   1.5     0.1   26.3 1997
## 1998 North America United States          3.4   1.2     0.2   26.2 1998
## 1999 North America United States          8.5   2.3     0.3    4.5 1999
## 2000          Asia         Japan          1.4  10.7     0.1    8.8 1999
##                                        Forbes.Webpage
## 1995          http://www.forbes.com/companies/shikoku-bank/
## 1996                http://www.forbes.com/companies/cameco/
## 1997 http://www.forbes.com/companies/bmce-banque-marocaine/
## 1998     http://www.forbes.com/companies/synovus-financial/
## 1999               http://www.forbes.com/companies/equifax/
## 2000    http://www.forbes.com/companies/uny-group-holdings/
```

c. Check the structure of **ForbesGlobal2000**.

Answer:

```
str(ForbesGlobal2000)
```

```
## 'data.frame':    2000 obs. of  11 variables:
##  $ Company       : chr  "ICBC" "China Construction Bank" "Agricultural Bank of China"
##  $ Sector        : chr  "Financials" "Financials" "Financials" "Financials" ...
##  $ Industry      : chr  "Major Banks" "Regional Banks" "Regional Banks" "Major Banks"
##  $ Continent     : chr  "Asia" "Asia" "Asia" "North America" ...
##  $ Country       : chr  "China" "China" "China" "United States" ...
##  $ Market.Value  : num  216 174 141 230 309 ...
##  $ Sales         : num  149 121 136 106 179 ...
##  $ Profits       : num  42.7 34.2 27 17.3 19.5 32.6 14.8 21.9 25.5 21.1 ...
##  $ Assets        : num  3125 2450 2405 2435 493 ...
##  $ Rank          : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ Forbes.Webpage: chr  "http://www.forbes.com/companies/icbc/" "http://www.forbes.co
```

d. Convert the variables **Sector** and **Industry** to factors.

Hint: you need to change the data type in the data frame, so the command should be like “`data-frame-name$column-name <-`”, and note that `attach()` cannot modify the data frame.

Check if there are any empty levels for the two factors. If so, replace those empty levels with `"NA"`. Then check your results again.

Hint: use the function `levels()` and see the link.

Answer:

```
ForbesGlobal2000$Sector   <- as.factor(ForbesGlobal2000$Sector)
ForbesGlobal2000$Industry <- as.factor(ForbesGlobal2000$Industry)

#To see if there is empty levels
levels(ForbesGlobal2000$Sector)
```

```
##  [1] ""                             "Consumer Discretionary"
##  [3] "Consumer Staples"             "Energy"
##  [5] "Financials"                   "Health Care"
##  [7] "Industrials"                  "Information Technology"
##  [9] "Materials"                    "Telecommunication Services"
## [11] "Utilities"
```

```
levels(ForbesGlobal2000$Industry)
```

```
##  [1] ""                                 "Advertising"
##  [3] "Aerospace & Defense"              "Air Courier"
##  [5] "Airline"                          "Aluminum"
##  [7] "Apparel/Accessories"              "Apparel/Footwear Retail"
##  [9] "Auto & Truck Manufacturers"       "Auto & Truck Parts"
## [11] "Beverages"                        "Biotechs"
## [13] "Broadcasting & Cable"             "Business & Personal Services"
## [15] "Business Products & Supplies"     "Casinos & Gaming"
## [17] "Communications Equipment"         "Computer & Electronics Retail"
## [19] "Computer Hardware"                "Computer Services"
## [21] "Computer Storage Devices"         "Conglomerates"
## [23] "Construction Materials"           "Construction Services"
## [25] "Consumer Electronics"             "Consumer Financial Services"
## [27] "Containers & Packaging"           "Department Stores"
## [29] "Discount Stores"                  "Diversified Chemicals"
## [31] "Diversified Insurance"            "Diversified Metals & Mining"
## [33] "Diversified Utilities"            "Drug Retail"
## [35] "Electric Utilities"               "Electrical Equipment"
## [37] "Electronics"                      "Environmental & Waste"
## [39] "Food Processing"                  "Food Retail"
## [41] "Furniture & Fixtures"             "Healthcare Services"
## [43] "Heavy Equipment"                  "Home Improvement Retail"
## [45] "Hotels & Motels"                  "Household Appliances"
## [47] "Household/Personal Care"          "Insurance Brokers"
## [49] "Internet & Catalog Retail"        "Investment Services"
## [51] "Iron & Steel"                     "Life & Health Insurance"
## [53] "Major Banks"                      "Managed Health Care"
## [55] "Medical Equipment & Supplies"     "Natural Gas Utilities"
## [57] "Oil & Gas Operations"           "Oil Services & Equipment"
## [59] "Other Industrial Equipment"     "Other Transportation"
## [61] "Paper & Paper Products"         "Pharmaceuticals"
## [63] "Precision Healthcare Equipment" "Printing & Publishing"
## [65] "Property & Casualty Insurance"  "Railroads"
## [67] "Real Estate"                    "Recreational Products"
## [69] "Regional Banks"                 "Rental & Leasing"
## [71] "Restaurants"                    "Security Systems"
## [73] "Semiconductors"                 "Software & Programming"
## [75] "Specialized Chemicals"          "Specialty Stores"
## [77] "Telecommunications services"    "Thrifts & Mortgage Finance"
## [79] "Tobacco"                        "Trading Companies"
## [81] "Trucking"
```

```r
levels(ForbesGlobal2000$Sector)[levels(ForbesGlobal2000$Sector) == ""] <- "NA"
levels(ForbesGlobal2000$Industry)[levels(ForbesGlobal2000$Industry) == ""] <- "NA"

#check again
levels(ForbesGlobal2000$Sector)
```

```
##  [1] "NA"                         "Consumer Discretionary"
##  [3] "Consumer Staples"           "Energy"
##  [5] "Financials"                 "Health Care"
##  [7] "Industrials"                "Information Technology"
##  [9] "Materials"                  "Telecommunication Services"
## [11] "Utilities"
```

```r
levels(ForbesGlobal2000$Industry)
```

```
##  [1] "NA"                             "Advertising"
##  [3] "Aerospace & Defense"            "Air Courier"
##  [5] "Airline"                        "Aluminum"
##  [7] "Apparel/Accessories"            "Apparel/Footwear Retail"
##  [9] "Auto & Truck Manufacturers"     "Auto & Truck Parts"
## [11] "Beverages"                      "Biotechs"
## [13] "Broadcasting & Cable"           "Business & Personal Services"
## [15] "Business Products & Supplies"   "Casinos & Gaming"
## [17] "Communications Equipment"       "Computer & Electronics Retail"
## [19] "Computer Hardware"              "Computer Services"
## [21] "Computer Storage Devices"       "Conglomerates"
## [23] "Construction Materials"         "Construction Services"
## [25] "Consumer Electronics"           "Consumer Financial Services"
## [27] "Containers & Packaging"         "Department Stores"
## [29] "Discount Stores"                "Diversified Chemicals"
## [31] "Diversified Insurance"          "Diversified Metals & Mining"
## [33] "Diversified Utilities"          "Drug Retail"
## [35] "Electric Utilities"                 "Electrical Equipment"
## [37] "Electronics"                        "Environmental & Waste"
## [39] "Food Processing"                    "Food Retail"
## [41] "Furniture & Fixtures"               "Healthcare Services"
## [43] "Heavy Equipment"                    "Home Improvement Retail"
## [45] "Hotels & Motels"                    "Household Appliances"
## [47] "Household/Personal Care"            "Insurance Brokers"
## [49] "Internet & Catalog Retail"          "Investment Services"
## [51] "Iron & Steel"                       "Life & Health Insurance"
## [53] "Major Banks"                        "Managed Health Care"
## [55] "Medical Equipment & Supplies"       "Natural Gas Utilities"
## [57] "Oil & Gas Operations"               "Oil Services & Equipment"
## [59] "Other Industrial Equipment"         "Other Transportation"
## [61] "Paper & Paper Products"             "Pharmaceuticals"
## [63] "Precision Healthcare Equipment" "Printing & Publishing"
## [65] "Property & Casualty Insurance"  "Railroads"
## [67] "Real Estate"                        "Recreational Products"
## [69] "Regional Banks"                     "Rental & Leasing"
## [71] "Restaurants"                        "Security Systems"
## [73] "Semiconductors"                     "Software & Programming"
## [75] "Specialized Chemicals"              "Specialty Stores"
## [77] "Telecommunications services"        "Thrifts & Mortgage Finance"
## [79] "Tobacco"                            "Trading Companies"
## [81] "Trucking"
```

“We can see the empty levels are replaced by NA”

e. Convert the variables **Continent** and **Country** to factors and check the number of levels (hint: use the function `nlevels()`).

Answer:

```
ForbesGlobal2000$Continent <- as.factor(ForbesGlobal2000$Continent)
ForbesGlobal2000$Country   <- as.factor(ForbesGlobal2000$Country)
nlevels(ForbesGlobal2000$Continent)
```

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

```
nlevels(ForbesGlobal2000$Country)
```

```
## [1] 63
```

“there are 6 Continent and 63 contries”

f. Sort the dataset by the continent and then by the country in alphabetical order. Print the first few observations of the sorted dataset.

Answer:

```
ForbesGlobal2000_sorted <- ForbesGlobal2000[order(ForbesGlobal2000$Continent,
                                                  ForbesGlobal2000$Country), ]
head(ForbesGlobal2000_sorted)
```

```
##                            Company     Sector               Industry Continent
## 1752 Commercial International Bank Financials         Regional Banks    Africa
## 1278                  Essar Energy     Energy   Oil & Gas Operations    Africa
## 1066             Attijariwafa Bank Financials         Regional Banks    Africa
## 1842     Banque Centrale Populaire Financials         Regional Banks    Africa
## 1997                     BMCE Bank Financials         Regional Banks    Africa
## 1046                Dangote Cement  Materials Construction Materials    Africa
##        Country Market.Value Sales Profits Assets Rank
## 1752     Egypt          4.8   1.7     0.4   16.4 1750
## 1278 Mauritius          1.5  27.8    -0.3   16.1 1278
## 1066   Morocco          7.9   2.8     0.5   44.9 1065
## 1842   Morocco          4.1   2.2     0.2   35.6 1842
## 1997   Morocco          4.6   1.5     0.1   26.3 1997
## 1046   Nigeria         24.3   2.4     1.3    5.3 1046
##                                                        Forbes.Webpage
## 1752 http://www.forbes.com/companies/commercial-international-bank/
## 1278                  http://www.forbes.com/companies/essar-energy/
## 1066             http://www.forbes.com/companies/attijariwafa-bank/
## 1842     http://www.forbes.com/companies/banque-centrale-populaire/
## 1997         http://www.forbes.com/companies/bmce-banque-marocaine/
## 1046                http://www.forbes.com/companies/dangote-cement/
```

g. Compute the mean and median of the profits of all companies from the first continent in part f.

Answer:

```
first_continent <- ForbesGlobal2000_sorted$Continent[1]
subset_first <- subset(ForbesGlobal2000_sorted, Continent == first_continent)
mean_profit   <- mean(subset_first$Profits, na.rm = TRUE)
median_profit <- median(subset_first$Profits, na.rm = TRUE)

mean_profit
```

```
## [1] 0.7538462
```

```
median_profit
```

```
## [1] 0.5
```

h. Create a subset of the dataset obtained in part f by excluding **Sales** and **Profits** and including only observations who are in the United States AND with **Assets** greater than 2000 billion USD. Print the first few observations of this dataset.

Answer:

```
subset_US <- subset(ForbesGlobal2000_sorted,
                    Country == "United States" & Assets > 2000,
                    select = -c(Sales, Profits))
head(subset_US)
```

```
##             Company     Sector            Industry     Continent       Country
## 4    JPMorgan Chase Financials         Major Banks North America United States
## 13  Bank of America Financials         Major Banks North America United States
## 355      Fannie Mae Financials Investment Services North America United States
##     Market.Value Assets Rank                                   Forbes.Webpage
## 4          229.7 2435.3    4  http://www.forbes.com/companies/jpmorgan-chase/
## 13         183.3 2113.8   13 http://www.forbes.com/companies/bank-of-america/
## 355          4.6 3270.1  355      http://www.forbes.com/companies/fannie-mae/
```

i. Create another subset of the dataset obtained in part f by including only **Profits**, **Assets**, and **Country**, and only observations who have **Profits** greater than 30 billion USD OR **Assets** greater than 2000 billion USD.

Answer:

```
subset_big <- subset(ForbesGlobal2000_sorted,
                     Profits > 30 | Assets > 2000,
                     select = c(Profits, Assets, Country))
head(subset_big)
```

```
##    Profits Assets Country
## 1     42.7 3124.9   China
## 2     34.2 2449.5   China
## 3     27.0 2405.4   China
## 9     25.5 2291.8   China
## 37    11.3 2458.9   Japan
## 21    39.0  397.2  Russia
```

j. Set the random seed to 124, and then randomly select 10 companies from the sorted dataset from part f and display their names and websites.

Answer:

```
set.seed(124)
random_index <- sample(1:nrow(ForbesGlobal2000_sorted), 10)
random_companies <- ForbesGlobal2000_sorted[random_index, c("Company", "Forbes.Webpage")
random_companies
```

```
##                         Company
## 604              Teck Resources
## 1880 Financial Street Holdings
## 339                   Accenture
## 117               Schlumberger
## 983                 Tata Steel
## 1169    Advanced Semiconductor
## 47                  Ford Motor
## 166            Delta Air Lines
## 1091                 PKN Orlen
## 1982                  Teradata
##                                                   Forbes.Webpage
## 604             http://www.forbes.com/companies/teck-resources/
## 1880 http://www.forbes.com/companies/financial-street-holdings/
## 339                   http://www.forbes.com/companies/accenture/
## 117                http://www.forbes.com/companies/schlumberger/
## 983                  http://www.forbes.com/companies/tata-steel/
## 1169    http://www.forbes.com/companies/advanced-semiconductor/
## 47                   http://www.forbes.com/companies/ford-motor/
## 166             http://www.forbes.com/companies/delta-air-lines/
## 1091                  http://www.forbes.com/companies/pkn-orlen/
## 1982                   http://www.forbes.com/companies/teradata/
```