Assignment 1 – Introduction to R

Total Points: 10

Instructions: Open the Assign1_Intro_R.Rproj R project file located within the Assign1_Intro_R subfolder of the Assign folder within 02_Data_Struct folder, and follow the instructions for each of the problems below. Unless otherwise indicated, each problem must be self-contained, i.e., if the R environment is cleared and only the code for a particular problem selected, the code must run without errors and produce the correct results to earn full credit for the problem. In addition, when indicated, you must use the vector, data frame, list and array names as specified in the problem. Failure to do so will result in an appropriate point deduction.

Problems 1 and 2 are located in Assign11_Vectors_DataFrames.r file.

Problem 1 (2 points): Nine States Population Density

The purpose of this problem is to calculate the population density and percentage of land area for the "top nine states" given in the various vectors at the top of the problem and in **Problems2-3** sheet of the Excel data file. The reason why these are referred as the "top nine states" is because these are the states with the largest number of major league sports associated with top 10 metro areas (California has 2) which will be used later in the assignment.

a) Calculate the population densities of each of the 9 states in **pop_density** vector, making sure these are rounded to 2 decimals. Then use the state abbreviations to create the names for each of the population density vector elements, and then display the vector showing the states with the highest density first and those with the lowest density last. The result is shown below:

```
     DC      MA      NY      FL      PA      CA      IL      TX      MN 
8848.87  835.16  410.43  348.64  283.43  238.87  230.83   96.05   66.62 
```

b) Using **tot_area** variable and **pct_area** vector, you must calculate and display the percent of land area each of the 9 states contributes to the total. These percentages, expressed on a 100 scale with 1 decimal, but without % symbols, must be displayed in decreasing order showing the state abbreviations on top of the numbers.

```
  TX   CA   MN   IL   FL   NY   PA   MA   DC 
37.0 22.1 11.3  7.9  7.6  6.7  6.3  1.1  0.0 
```

Problem 2 (2 points): Nine States Population Change

This problem will examine the population change between 2010 and 2020 for the same nine states.

a) Create the **top_states_pop_df** data frame using its constituent vectors. Create the **tsp_names** vector with **Name**, **Abbr**, **Area**, **Pop2010** and **Pop2020**, and used it to modify the default data frame names. Then display the entire data frame and run the **summary** function on it. The result should look like this:

```
                  Name Abbr   Area  Pop2010  Pop2020
1             New York   NY  47214 19378102 20201249
2           California   CA 155959 37253956 39538223
3             Illinois   IL  55584 12830632 12821508
4                Texas   TX 261797 25145561 29145505
5 District of Columbia   DC     68   601723   689545
6         Pennsylvania   PA  44817 12702379 13002700
7              Florida   FL  53927 18801310 21538187
8        Massachusetts   MA   7840  6547629  7029917
9            Minnesota   MN  79610  5303925  5706494
```

```
     Name               Abbr                Area           Pop2010            Pop2020        
 Length:9           Length:9           Min.   :    68   Min.   :  601723   Min.   :  689545  
 Class :character   Class :character   1st Qu.: 44817   1st Qu.: 6547629   1st Qu.: 7029917  
 Mode  :character   Mode  :character   Median : 53927   Median :12830632   Median :13002700  
                                       Mean   : 78535   Mean   :15396135   Mean   :16630370  
                                       3rd Qu.: 79610   3rd Qu.:19378102   3rd Qu.:21538187  
                                       Max.   :261797   Max.   :37253956   Max.   :39538223  
```

b) Add the population percentage change, on a 100 scale, rounded to 2 decimals (but without % symbols), to the top_states_pop_df data frame as a **PopPctChg** column and display the entire data frame). You must use the data frame columns, instead of the data vectors, using the $ operator with defined names.

```
                  Name Abbr   Area  Pop2010  Pop2020 PopPctChg
1             New York   NY  47214 19378102 20201249      4.25
2           California   CA 155959 37253956 39538223      6.13
3             Illinois   IL  55584 12830632 12821508     -0.07
4                Texas   TX 261797 25145561 29145505     15.91
5 District of Columbia   DC     68   601723   689545     14.60
6         Pennsylvania   PA  44817 12702379 13002700      2.36
7              Florida   FL  53927 18801310 21538187     14.56
8        Massachusetts   MA   7840  6547629  7029917      7.37
9            Minnesota   MN  79610  5303925  5706494      7.59
```

Problems 3 and 4 are located in `Assign12_Lists_Arrays.r` file.

Problem 3 (3 points): Top 10 Metro Sports

Before you begin with this problem, review the links to the major league sports data on Wikipedia:

- https://en.wikipedia.org/wiki/List_of_American_and_Canadian_cities_by_number_of_major_professional_sports_franchises

and Nielsen's designated market area (DMA) rankings:

- https://mediatracks.com/resources/nielsen-dma-rankings-2020/.

The key thing is to make sure you understand the way the (no longer up to date) data is represented in R, starting from the top of the file to about line 80. Specifically, you need to review the 10 lists of sports for each of the top 10 metro areas. Each element of each list is a 2-element vector, with the first vector element listing the major league sport and the second the name of the team.

Before you continue with the rest of the problem, **make sure to run the first 80 lines of code!!!**

a) Create and display **Metro_Stats_DF** data frame with the 5 vectors from the top of the file. These data vectors contain the non-sports data, from the name of the city, abbreviation, region, 2016 population estimate and Nielsen's estimate for the number of TV homes. Review the entire data frame and make sure to display the last MSP row. You must use the referencing that will produce the output below.

```
                 CityMetro MetroAbbr Region PopEst2016 TV_Homes
10 Minneapolis-Saint Paul        MSP    MDW    3551036  1697370
```

b) Then you need to create a master list of all the sports in all top 10 metro areas. This master-list consists of 10 individual lists already provided. In other words, the **ML_Sports_List** contains each of the individual metro sports lists as its elements. You should display both the data frame and the list, so you can check the list structure specifically. Review the entire master-list and make sure to display the name of the local (MSP) hockey team. You must use the referencing that will produce the output below.

```
   Sport     Team
"Hockey"   "Wild"
```

Using a for loop, run the uber-master list of all provided non-sports info and the master list of all major league sports for the top 10 metro areas. To make it more precise, the **i-th** element of this uber-master list, named **Metro_ML_Sports_List**, consists of the **i-th** row in the Metro_Stats_DF data frame with **Stats** for the name of this element. The second element, named **Sports**, is the **i-th** ML_Sports_List element containing the list of sports for that metro area.

c) Review the entire uber-list and then display the MSP's 2016 population estimate and number of TV homes, followed by the local (MSP) hockey team. You must use the referencing that will produce the output below. There are two separate lines of code producing the result below.

```
   PopEst2016 TV_Homes
10    3551036  1697370


   Sport     Team
"Hockey"   "Wild"
```

Problem 4 (3 points): Top 10 Metro Array

The objective of this problem is to recognize how the 3D array of population breakdowns by age group in each of the 10 metro areas and for 2 different years, 2008 and 1996. The already provided code includes the labels for the metro abbreviations, each of the 3 age groups, and the 2 years. The data is already entered by showing the 2008 percentage population in the younger than 35 group for all 10 metro areas. This is followed by the 2008 population breakdown for the middle age group from 35 to 64, followed by the oldest age group cohort breakdowns in the same year. The corresponding numbers in the first 3 rows roughly add to 100, but not exactly due to rounding errors. The whole process is then repeated for year 1996. Finally, the dimensions of 10 metros, 3 age groups and 2 years round up the definition of the array. Using the provided code, run the 3D **metro_age_array** and use it to answer the questions below.

a) Display only the data for the MSP metro. You must use the referencing that will produce the output below.

```
          2008 1996
Yng34     47.9 52.6
Mid35_64  42.0 37.7
Old65     10.2  9.7
```

b) Calculate the array of percent population differences for each metro area and age group between 2008 and 1996. A negative number represents the relative decrease in population for that age group in the particular metro area. The positive number represents the relative increase.

```
    Yng34 Mid35_64 Old65
NYC  -3.1      3.5  -0.3
LA   -5.9      5.5   0.7
CHI  -3.3      3.2  -0.2
SF   -6.6      5.7   0.9
DFW  -2.2      2.3  -0.2
DC   -3.5      3.8  -0.2
PHL  -3.7      4.3  -0.6
MFL  -4.5      4.6  -0.2
BOS  -4.7      5.2  -0.5
MSP  -4.7      4.3   0.5
```

c) Find the highest percent increase and the lowest percent decrease in the array. Save the results in the **high_pct** and **low_pct** variables and display their contents shown below. Because the array is small, you can manually identify the metro and age group both the highest and lowest percentage changes belong to. There are two separate lines of code producing the result below.

```
[1] 5.7
[1] -6.6
```

Submission: You must submit the completed `Assign1_Intro_R` project folder, zipped up into a compressed folder by the same name on Canvas by the designated due date.