INTRODUCTION TO DATA SCIENCE
R Programming Exercises

Member Names:
Alira Ilagan **Date:**_____
Larry Miguel Cueva **Year and Section: BSCS 4-2**

College Grades Data

Remember the college student's data? We'll compute their final grades.

The grade computation will be:

- **30%:** Average grade of long exams (2 long exams)
- **40%:** Final exam grade (1 final exam)
- **30%:** HW/Machine Problems (3 Homeworks)

Class policy

Before the course started, you also have the discussed to them the following policies:

1. The grading scale is:

Final Grade	Final Grade in Percentage
1.0	[80 – 100]
2.0	[70 – 79]
3.0	[50 – 69]
4.0	[0 – 49]

2. Max absences is 6. Students who exceeded the limit will get a grade of 5.0.
3. Missed exams or HW: 0
4. Cheating will get 0 points on that specific HW/Exam.

Additional Premise

1. You will see that almost everyone cheated in `HW 3`, except for **Chrisphee**. They all have the same answers and got `97`. Change the `HW 3` scores of students who cheated to `0`, and change Chrisphee's HW score to `100`.
2. You promised a bonus of `5` points to be added in their final percentage grades.

1. Data Audit

1.1. Import library 'dplyr' and load College Students dataset in R

Load the **college_students_w_header.csv** data here and store it as **college_df.** Check the data frame then see the expected output below.

```
library(dplyr)

college_df <- read.csv(file.choose(), header=T)

college_df
```

Expected Output:

	Surname	First.Name	Section	Sex	Student.Number	Absences	HW.1	HW.2	HW.3	Exam.1	Exam.2	Final.Exam
1	Anderson	Totoy	A	M	2007-59081	8	59	48	97	74	74	77
2	Smith	Neneng	A	F	2011-39123	3	55	67	97	78	92	84
3	Reyes	Reginald	B	M	2014-19283	2	52	63	97	75	85	83
4	Cruz	Conrad	A	M	2014-12345	0	58	69	97	95	89	92
5	Dy	Michelle	B	F	2014-23456	4	56	60	97	76	81	80
6	Sy	Enrique	A	M	2014-34905	3	47	NA	97	79	85	86
7	Mariano	Mark	B	M	2014-00501	1	65	59	97	81	89	80
8	Mulimbayan	Benedict	B	M	2013-86942	1	53	55	97	91	91	85
9	Sung	Sam	B	M	2008-19492	3	55	55	97	78	91	81
10	Bacon	Chrisphee	A	M	2008-23748	3	57	64	60	86	NA	91
11	Daniels	Jack	A	M	2014- 50123	1	69	61	97	87	81	94
12	Ball	Krystall	B	F	2015-10301	0	48	59	97	80	82	86
13	Torres	Erica	A	F	2014-00239	6	52	54	97	72	82	79
14	Salamat	Andrea	B	F	2013-93816	1	55	55	97	79	88	80
15	Gomez	Chelsea	A	F	2014-11111	5	61	55	97	78	84	85
16	Reyes	Angelica	A	F	2014-87908	0	58	66	97	90	80	90
17	Reyes	Phoebe	B	F	2014-54656	0	57	55	97	86	83	83
18	Reyes	Zach	B	M	2013-47383	2	51	59	97	84	86	82
19	Torres	Veronika	A	F	2013-44343	1	62	73	97	87	90	85
20	Reyes	Anna	B	F	2014-14314	7	50	54	97	76	78	78
21	<NA>	<NA>	<NA>	<NA>	<NA>	NA	NA	NA	NA	NA	NA	NA

1.2. Check how many entries do we have?

```
cat("\"the number of entries are: \"", nrow(college_df))
```

Expected Output:

```
[1] "The number of entries are :  21"
```

1.3. Remove the entries that have no "Surname" and "First Name"

```
college_df <- college_df[!is.na(college_df$Surname) & !is.na(college_df$First.Name),]
```

Expected Output:

	Surname	First.Name	Section	Sex	Student.Number	Absences	HW.1	HW.2	HW.3	Exam.1	Exam.2	Final.Exam
15	Gomez	Chelsea	A	F	2014-11111	5	61	55	97	78	84	85
16	Reyes	Angelica	A	F	2014-87908	0	58	66	97	90	80	90
17	Reyes	Phoebe	B	F	2014-54656	0	57	55	97	86	83	83
18	Reyes	Zach	B	M	2013-47383	2	51	59	97	84	86	82
19	Torres	Veronika	A	F	2013-44343	1	62	73	97	87	90	85
20	Reyes	Anna	B	F	2014-14314	7	50	54	97	76	78	78

1.4. How many students did not take the “HW 2”?

```
nrow(college_df[is.na(college_df$HW.2),])
```

Expected Output:

```
[1] 1
```

1.5. Who are they?

```
hw2_absentee <- college_df[is.na(college_df$HW.2),]
select(hw2_absentee, Surname, First.Name)
```

Expected Output:

```
  Surname First.Name
6      Sy    Enrique
```

1.6. How many students did not take the “Exam 2”?

```
nrow(college_df[is.na(college_df$Exam.2),])
```

Expected Output:

```
[1] 1
```

1.7. Who are they?

```
exam2_absentee <- college_df[is.na(college_df$Exam.2),]
select(exam2_absentee, Surname, First.Name)
```

Expected Output:

```
   Surname First.Name
10   Bacon  Chrisphee
```

2. Data Cleaning and Processing

Checklist:

- **Add new columns**
  - “***Full Name***” from “**Surname**” and “**First Name**”, please follow the format *<first_name> <surname>*, example: Chrisphee Bacon
- **Resolve the cheating incident**
  - HW 3 grades should be 0
  - Chrisphee's HW 3 grade should be 100 because of her honestly.
- **Resolve the missed HWs and Exams**
  - HW 2
  - Exam 2

2.1. Add new columns – "Full Name"

```
full_name <- paste(college_df$First.Name, college_df$Surname)
college_df['Full name'] <- full_name
```

Expected Output:

	Surname	First.Name	Section	Sex	Student.Number	Absences	HW.1	HW.2	HW.3	Exam.1	Exam.2	Final.Exam	Full Name
1	Anderson	Totoy	A	M	2007-59081	8	59	48	97	74	74	77	Totoy Anderson
2	Smith	Neneng	A	F	2011-39123	3	55	67	97	78	92	84	Neneng Smith
3	Reyes	Reginald	B	M	2014-19283	2	52	63	97	75	85	83	Reginald Reyes
4	Cruz	Conrad	A	M	2014-12345	0	58	69	97	95	89	92	Conrad Cruz
5	Dy	Michelle	B	F	2014-23456	4	56	60	97	76	81	80	Michelle Dy
6	Sy	Enrique	A	M	2014-34905	3	47	NA	97	79	85	86	Enrique Sy

2.2. Perform changes based on the cheating incident.

```
cheaters_to_zero <- function(hw3_x){
 print(hw3_x)
 return (if(hw3_x >= 97) 0 else 100)
}

college_df$HW.3 <- sapply(college_df$HW.3, cheaters_to_zero)
print(college_df)
```

Expected Output:

	Surname	First.Name	Section	Sex	Student.Number	Absences	HW.1	HW.2	HW.3	Exam.1	Exam.2	Final.Exam	Full Name
1	Anderson	Totoy	A	M	2007-59081	8	59	48	0	74	74	77	Totoy Anderson
2	Smith	Neneng	A	F	2011-39123	3	55	67	0	78	92	84	Neneng Smith
3	Reyes	Reginald	B	M	2014-19283	2	52	63	0	75	85	83	Reginald Reyes
4	Cruz	Conrad	A	M	2014-12345	0	58	69	0	95	89	92	Conrad Cruz
5	Dy	Michelle	B	F	2014-23456	4	56	60	0	76	81	80	Michelle Dy
6	Sy	Enrique	A	M	2014-34905	3	47	NA	0	79	85	86	Enrique Sy

2.3. Chrisphee's HW 3 grade should be 100 because of her honesty.

```
cheaters_to_zero <- function(hw3_x){
 print(hw3_x)
 return (if(hw3_x >= 97) 0 else 100)
}

college_df$HW.3 <- sapply(college_df$HW.3, cheaters_to_zero)
print(college_df)
```

Expected Output:

	Surname	First.Name	Section	Sex	Student.Number	Absences	HW.1	HW.2	HW.3	Exam.1	Exam.2	Final.Exam	Full Name
1	Anderson	Totoy	A	M	2007-59081	8	59	48	0	74	74	77	Totoy Anderson
2	Smith	Neneng	A	F	2011-39123	3	55	67	0	78	92	84	Neneng Smith
3	Reyes	Reginald	B	M	2014-19283	2	52	63	0	75	85	83	Reginald Reyes
4	Cruz	Conrad	A	M	2014-12345	0	58	69	0	95	89	92	Conrad Cruz
5	Dy	Michelle	B	F	2014-23456	4	56	60	0	76	81	80	Michelle Dy
6	Sy	Enrique	A	M	2014-34905	3	47	NA	0	79	85	86	Enrique Sy
7	Mariano	Mark	B	M	2014-00501	1	65	59	0	81	89	80	Mark Mariano
8	Mulimbayan	Benedict	B	M	2013-86942	1	53	55	0	91	91	85	Benedict Mulimbayan
9	Sung	Sam	B	M	2008-19492	3	55	55	0	78	91	81	Sam Sung
10	Bacon	Chrisphee	A	M	2008-23748	3	57	64	100	86	NA	91	Chrisphee Bacon
11	Daniels	Jack	A	M	2014- 50123	1	69	61	0	87	81	94	Jack Daniels
12	Ball	Krystall	B	F	2015-10301	0	48	59	0	80	82	86	Krystall Ball
13	Torres	Erica	A	F	2014-00239	6	52	54	0	72	82	79	Erica Torres
14	Salamat	Andrea	B	F	2013-93816	1	55	55	0	79	88	80	Andrea Salamat
15	Gomez	Chelsea	A	F	2014-11111	5	61	55	0	78	84	85	Chelsea Gomez
16	Reyes	Angelica	A	F	2014-87908	0	58	66	0	90	80	90	Angelica Reyes
17	Reyes	Phoebe	B	F	2014-54656	0	57	55	0	86	83	83	Phoebe Reyes
18	Reyes	Zach	B	M	2013-47383	2	51	59	0	84	86	82	Zach Reyes

2.4. Resolve missed HW's and Exams.

```
college_df[is.na(college_df)] <- "0"
```

Expected Output:

	Surname	First.Name	Section	Sex	Student.Number	Absences	HW.1	HW.2	HW.3	Exam.1	Exam.2	Final.Exam	Full Name
1	Anderson	Totoy	A	M	2007-59081	8	59	48	0	74	74	77	Totoy Anderson
2	Smith	Neneng	A	F	2011-39123	3	55	67	0	78	92	84	Neneng Smith
3	Reyes	Reginald	B	M	2014-19283	2	52	63	0	75	85	83	Reginald Reyes
4	Cruz	Conrad	A	M	2014-12345	0	58	69	0	95	89	92	Conrad Cruz
5	Dy	Michelle	B	F	2014-23456	4	56	60	0	76	81	80	Michelle Dy
6	Sy	Enrique	A	M	2014-34905	3	47	0	0	79	85	86	Enrique Sy
7	Mariano	Mark	B	M	2014-00501	1	65	59	0	81	89	80	Mark Mariano
8	Mulimbayan	Benedict	B	M	2013-86942	1	53	55	0	91	91	85	Benedict Mulimbayan
9	Sung	Sam	B	M	2008-19492	3	55	55	0	78	91	81	Sam Sung
10	Bacon	Chrisphee	A	M	2008-23748	3	57	64	100	86	0	91	Chrisphee Bacon
11	Daniels	Jack	A	M	2014- 50123	1	69	61	0	87	81	94	Jack Daniels
12	Ball	Krystall	B	F	2015-10301	0	48	59	0	80	82	86	Krystall Ball
13	Torres	Erica	A	F	2014-00239	6	52	54	0	72	82	79	Erica Torres
14	Salamat	Andrea	B	F	2013-93816	1	55	55	0	79	88	80	Andrea Salamat
15	Gomez	Chelsea	A	F	2014-11111	5	61	55	0	78	84	85	Chelsea Gomez
16	Reyes	Angelica	A	F	2014-87908	0	58	66	0	90	80	90	Angelica Reyes
17	Reyes	Phoebe	B	F	2014-54656	0	57	55	0	86	83	83	Phoebe Reyes
18	Reyes	Zach	B	M	2013-47383	2	51	59	0	84	86	82	Zach Reyes
19	Torres	Veronika	A	F	2013-44343	1	62	73	0	87	90	85	Veronika Torres
20	Reyes	Anna	B	F	2014-14314	7	50	54	0	76	78	78	Anna Reyes

2.5. Computation of Grades

Checklist:

- Compute for the final grades. Don't forget to add **5** points in their Final Grade (**%**), you kind-hearted person you.
- Convert their grades based on the scale.
- Set the grade to **5.0** for students with excessive absences.

2.5.1. Get the average mean of the HWs and Exams.

Before getting the **"HW mean"** and **"Exam mean",** check first if these attributes: **'HW.1'**, **'HW.2'**, **'HW.3'**, **'Exam.1'**, and **'Exam.2'** are in numeric or integer data type.

typeof(college_df$HW.1)
typeof(college_df$HW.2)
typeof(college_df$HW.3)
typeof(college_df$Exam.1)
typeof(college_df$Exam.2)

```
> typeof(college_df$HW.1)
[1] "integer"
> typeof(college_df$HW.2)
[1] "character"
> typeof(college_df$HW.3)
[1] "character"
> typeof(college_df$Exam.1)
[1] "integer"
> typeof(college_df$Exam.2)
[1] "character"
```

Just in case, if one of them are in a ***character data type***, this should be converted into **numeric data type**.

```
college_df$HW.2 = as.integer(college_df$HW.2)
college_df$HW.3 = as.integer(college_df$HW.3)
college_df$Exam.2 = as.integer(college_df$Exam.2)
```

Get the **mean** and store them in columns **"HW Mean"** and **"Exam Mean"**, respectively.

```
college_df$HW.Mean <- rowMeans(subset(college_df, select = c(HW.1, HW.2, HW.3)), )
college_df$Exam.Mean <- rowMeans(subset(college_df, select = c(Exam.1, Exam.2)), )
```

Expected Output:

	Surname	First.Name	Section	Sex	Student.Number	Absences	HW.1	HW.2	HW.3	Exam.1	Exam.2	Final.Exam	Full Name	HW Mean	Exam Mean
1	Anderson	Totoy	A	M	2007-59081	8	59	48	0	74	74	77	Totoy Anderson	35.66667	74.0
2	Smith	Neneng	A	F	2011-39123	3	55	67	0	78	92	84	Neneng Smith	40.66667	85.0
3	Reyes	Reginald	B	M	2014-19283	2	52	63	0	75	85	83	Reginald Reyes	38.33333	80.0
4	Cruz	Conrad	A	M	2014-12345	0	58	69	0	95	89	92	Conrad Cruz	42.33333	92.0
5	Dy	Michelle	B	F	2014-23456	4	56	60	0	76	81	80	Michelle Dy	38.66667	78.5
6	Sy	Enrique	A	M	2014-34905	3	47	0	0	79	85	86	Enrique Sy	15.66667	82.0

2.5.2. Compute Final Grade in Percentage (FG Per)

Use the columns **Final Exam**, **HW Mean**, and **Exam Mean**.

For reference:

- **30%**: Mean grade of long exams (2 long exams)
- **40%**: Final exam grade (1 final exam)
- **30%**: HW/Machine Problems (3 homeworks)

```
college_df$FG.Per <- (college_df$Exam.Mean* .3 + college_df$Final.Exam* .4 +
college_df$HW.Mean* .3) + 5
```

Expected Output:

	Surname	First.Name	Section	Sex	Student.Number	Absences	HW.1	HW.2	HW.3	Exam.1	Exam.2	Final.Exam	Full Name	HW Mean	Exam Mean	FG Per
1	Anderson	Totoy	A	M	2007-59081	8	59	48	0	74	74	77	Totoy Anderson	35.66667	74.0	68.70
2	Smith	Neneng	A	F	2011-39123	3	55	67	0	78	92	84	Neneng Smith	40.66667	85.0	76.30
3	Reyes	Reginald	B	M	2014-19283	2	52	63	0	75	85	83	Reginald Reyes	38.33333	80.0	73.70
4	Cruz	Conrad	A	M	2014-12345	0	58	69	0	95	89	92	Conrad Cruz	42.33333	92.0	82.10
5	Dy	Michelle	B	F	2014-23456	4	56	60	0	76	81	80	Michelle Dy	38.66667	78.5	72.15
6	Sy	Enrique	A	M	2014-34905	3	47	0	0	79	85	86	Enrique Sy	15.66667	82.0	68.70

2.5.3. Get the scaled grades

This should be based from the **FG Per** column.

For reference:

Final Grade	Final Grade in Percentage
1.0	[80 – 100]
2.0	[70 – 79]
3.0	[50 – 69]
4.0	[0 – 49]

```
college_df$Final.Grade <- ifelse((between(college_df$FG.Per, 80 , 100)), 1,
ifelse((between(college_df$FG.Per, 70, 79.99)), 2,  ifelse((between(college_df$FG.Per, 50, 69.99)),
3, ifelse((between(college_df$FG.Per, 0, 49.99)), 4, "")))))
```

Expected Output:

	Surname	First.Name	Section	Sex	Student.Number	Absences	HW.1	HW.2	HW.3	Exam.1	Exam.2	Final.Exam	Full Name	HW Mean	Exam Mean	FG Per	Final Grade
1	Anderson	Totoy	A	M	2007-59081	8	59	48	0	74	74	77	Totoy Anderson	35.66667	74.0	68.70	3
2	Smith	Neneng	A	F	2011-39123	3	55	67	0	78	92	84	Neneng Smith	40.66667	85.0	76.30	2
3	Reyes	Reginald	B	M	2014-19283	2	52	63	0	75	85	83	Reginald Reyes	38.33333	80.0	73.70	2
4	Cruz	Conrad	A	M	2014-12345	0	58	69	0	95	89	92	Conrad Cruz	42.33333	92.0	82.10	1
5	Dy	Michelle	B	F	2014-23456	4	56	60	0	76	81	80	Michelle Dy	38.66667	78.5	72.15	2
6	Sy	Enrique	A	M	2014-34905	3	47	0	0	79	85	86	Enrique Sy	15.66667	82.0	68.70	3
7	Mariano	Mark	B	M	2014-00501	1	65	59	0	81	89	80	Mark Mariano	41.33333	85.0	74.90	2
8	Mulimbayan	Benedict	B	M	2013-86942	1	53	55	0	91	91	85	Benedict Mulimbayan	36.00000	91.0	77.10	2
9	Sung	Sam	B	M	2008-19492	3	55	55	0	78	91	81	Sam Sung	36.66667	84.5	73.75	2
10	Bacon	Chrisphee	A	M	2008-23748	3	57	64	100	86	0	91	Chrisphee Bacon	73.66667	43.0	76.40	2
11	Daniels	Jack	A	M	2014- 50123	1	69	61	0	87	81	94	Jack Daniels	43.33333	84.0	80.80	1
12	Ball	Krystall	B	F	2015-10301	0	48	59	0	80	82	86	Krystall Ball	35.66667	81.0	74.40	2
13	Torres	Erica	A	F	2014-00239	6	52	54	0	72	82	79	Erica Torres	35.33333	77.0	70.30	2
14	Salamat	Andrea	B	F	2013-93816	1	55	55	0	79	88	80	Andrea Salamat	36.66667	83.5	73.05	2
15	Gomez	Chelsea	A	F	2014-11111	5	61	55	0	78	84	85	Chelsea Gomez	38.66667	81.0	74.90	2
16	Reyes	Angelica	A	F	2014-87908	0	58	66	0	90	80	90	Angelica Reyes	41.33333	85.0	78.90	2
17	Reyes	Phoebe	B	F	2014-54656	0	57	55	0	86	83	83	Phoebe Reyes	37.33333	84.5	74.75	2
18	Reyes	Zach	B	M	2013-47383	2	51	59	0	84	86	82	Zach Reyes	36.66667	85.0	74.30	2
19	Torres	Veronika	A	F	2013-44343	1	62	73	0	87	90	85	Veronika Torres	45.00000	88.5	79.05	2
20	Reyes	Anna	B	F	2014-14314	7	50	54	0	76	78	78	Anna Reyes	34.66667	77.0	69.70	3

2.5.4. Resolve excessive absences

```
college_df$Final.Grade[college_df$Absences > 6] <- 5
```

	Surname	First.Name	Section	Sex	Student.Number	Absences	HW.1	HW.2	HW.3	Exam.1	Exam.2	Final.Exam	Full Name	HW Mean	Exam Mean	FG Per	Final Grade
1	Anderson	Totoy	A	M	2007-59081	8	59	48	0	74	74	77	Totoy Anderson	35.66667	74.0	68.70	5
2	Smith	Neneng	A	F	2011-39123	3	55	67	0	78	92	84	Neneng Smith	40.66667	85.0	76.30	2
3	Reyes	Reginald	B	M	2014-19283	2	52	63	0	75	85	83	Reginald Reyes	38.33333	80.0	73.70	2
4	Cruz	Conrad	A	M	2014-12345	0	58	69	0	95	89	92	Conrad Cruz	42.33333	92.0	82.10	1
5	Dy	Michelle	B	F	2014-23456	4	56	60	0	76	81	80	Michelle Dy	38.66667	78.5	72.15	2

FINAL OUTPUT:

	Surname	First.Name	Section	Sex	Student.Number	Absences	HW.1	HW.2	HW.3	Exam.1	Exam.2	Final.Exam	Full Name	HW Mean	Exam Mean	FG Per	Final Grade
1	Anderson	Totoy	A	M	2007-59081	8	59	48	0	74	74	77	Totoy Anderson	35.66667	74.0	68.70	5
2	Smith	Neneng	A	F	2011-39123	3	55	67	0	78	92	84	Neneng Smith	40.66667	85.0	76.30	2
3	Reyes	Reginald	B	M	2014-19283	2	52	63	0	75	85	83	Reginald Reyes	38.33333	80.0	73.70	2
4	Cruz	Conrad	A	M	2014-12345	0	58	69	0	95	89	92	Conrad Cruz	42.33333	92.0	82.10	1
5	Dy	Michelle	B	F	2014-23456	4	56	60	0	76	81	80	Michelle Dy	38.66667	78.5	72.15	2
6	Sy	Enrique	A	M	2014-34905	3	47	0	0	79	85	86	Enrique Sy	15.66667	82.0	68.70	3
7	Mariano	Mark	B	M	2014-00501	1	65	59	0	81	89	80	Mark Mariano	41.33333	85.0	74.90	2
8	Mulimbayan	Benedict	B	M	2013-86942	1	53	55	0	91	91	85	Benedict Mulimbayan	36.00000	91.0	77.10	2
9	Sung	Sam	B	M	2008-19492	3	55	55	0	78	91	81	Sam Sung	36.66667	84.5	73.75	2
10	Bacon	Chrisphee	A	M	2008-23748	3	57	64	100	86	0	91	Chrisphee Bacon	73.66667	43.0	76.40	2
11	Daniels	Jack	A	M	2014- 50123	1	69	61	0	87	81	94	Jack Daniels	43.33333	84.0	80.80	1
12	Ball	Krystall	B	F	2015-10301	0	48	59	0	80	82	86	Krystall Ball	35.66667	81.0	74.40	2
13	Torres	Erica	A	F	2014-00239	6	52	54	0	72	82	79	Erica Torres	35.33333	77.0	70.30	2
14	Salamat	Andrea	B	F	2013-93816	1	55	55	0	79	88	80	Andrea Salamat	36.66667	83.5	73.05	2
15	Gomez	Chelsea	A	F	2014-11111	5	61	55	0	78	84	85	Chelsea Gomez	38.66667	81.0	74.90	2
16	Reyes	Angelica	A	F	2014-87908	0	58	66	0	90	80	90	Angelica Reyes	41.33333	85.0	78.90	2
17	Reyes	Phoebe	B	F	2014-54656	0	57	55	0	86	83	83	Phoebe Reyes	37.33333	84.5	74.75	2
18	Reyes	Zach	B	M	2013-47383	2	51	59	0	84	86	82	Zach Reyes	36.66667	85.0	74.30	2
19	Torres	Veronika	A	F	2013-44343	1	62	73	0	87	90	85	Veronika Torres	45.00000	88.5	79.05	2
20	Reyes	Anna	B	F	2014-14314	7	50	54	0	76	78	78	Anna Reyes	34.66667	77.0	69.70	5

End of Exercise.