Assignment 3: Data Cleaning

One of the big issues when it comes to working with data in any context is the issue of **data cleaning, Wrangling and merging of datasets**, since it is often the case that you will find yourself having to collate data across multiple files, and will need to rely on R to carry out

1. **sales:** This file contains the variables **Date**, **ID** (which is Product ID), and **Sales**. Load this into R under the name **mydata**.

2. **customers:** This file contains the variables **ID**, **Age**, and **Country**. Load this into R under the name **mydata2**.

Do the following 6 tasks in R using these two data sets

1. Storing variables in a data frame

To start off with a simple example, let us choose the customers dataset. Suppose that we only wish to include the variables ID and Age in a new data frame named *dataframe*

2 using the merge functions

Often times, it is necessary to combine two variables from different datasets to join two variables based on certain criteria. In R, this can be done using the **merge** function.

For instance, suppose that we wish to link the **Date** variable in the sales dataset with the **Age** and **Country** variables in the customers dataset – with the **ID** variable being the common link.

Upon doing this, we see that a new dataset is formed in R joining our chosen variables:

	ID	Sales	Name	Age	Country
13	13	125303	Janella Landrum	45	American Samoa
18	18	116634	Venetta Amante	23	American Samoa
30	30	157918	Jesusa Divers	50	Angola
77	77	56288	Callie Nilsen	44	Anguilla
85	85	96601	Mayme Nordstrom	41	Antigua and Barbuda
58	58	135896	Kara Creek	23	Belarus
59	59	138120	Ashly Yelverton	47	Belize
19	19	149661	Bettina Agee	26	Bolivia
93	93	100399	Yesenia Hugh	31	Bolivia

Showing 1 to 9 of 100 entries

3. Using as.date to format dates and calculate duration

Suppose that we now wish to calculate the number of days between the current date and the date of sale as listed in the sales file. In order to accomplish this, we can use as.date

Going back to the example above, suppose that we now wish to combine this duration variable with the rest of our data.

Hence, we can now combine our new **Duration** variable with the *merge* function

	ID	Sales	Date	Duration
1	48	113769	2014-02-12	1037
2	51	122965	2014-02-14	1035
3	4	164556	2014-03-18	1003
4	90	178351	2014-03-30	991
5	32	158446	2014-04-09	981
6	72	130730	2014-04-09	981
7	74	135108	2014-04-11	979
8	16	149196	2014-05-04	956
9	3	171482	2014-05-08	952
10	59	116634	2014-05-09	951
11	99	169763	2014-05-12	948
12	92	134180	2014-05-13	947
13	71	109975	2014-05-30	930

4. grepl: Remove instances of a string from a data

Let us look to the Country variable. We wish to remove all instances of “Greenland” from our customer data frame. This is accomplished using the ***grepl*** command:

5. Delete observations using head and tail functions

The head and tail functions can be used if we wish to delete certain observations from a variable, e.g. Sales. The head function allows us to delete the first 30 rows, while the tail function allows us to delete the last 30 rows. Do it using R and convert the new data into two matrices using as.matrix function

6 Using the "aggregate" function

Create a column names having these values *"John", "Elizabeth", "Michael", "John", "Elizabeth", "Michael"*
create a column webvisitsframe using these values "24","32","40","71","65","63"
convert webvisits in numeric data

Create a column minutesspentframe having these values "20", "41", "5", "6", "48", "97"
convert minutesspent into numeric format

The result is given in the following table now we want to obtain the sum of web visits and minutes spent on a website in any particular period:

	names	webvisits	minutesspent
1	John	24	20
2	Elizabeth	32	41
3	Michael	40	5
4	John	71	6
5	Elizabeth	65	48
6	Michael	63	97

In this instance, we can have to sum all webvisits by names using aggregate function
Thus, the values associated with identifiers (in this case, names) are summed up as follows:

	names	webvisits	minutesspent
1	Elizabeth	97	89
2	John	95	26
3	Michael	103	102