How to clean the datasets in R?, Data cleansing is one of the important steps in data analysis. Multiple packages are available in r to clean the data sets, here we are going to explore the janitor package to examine and clean the data.

Data cleaning is the process of transforming dirty data into reliable data that can be analyzed. Data cleansing improves your data quality and overall productivity.

When you clean your data, all incorrect information is gone and leaving only reliable quality information.

The main functions of the Janitor package are

- Format ugly data frame column names
- Isolate duplicate records in the data frame
- Provide quick tabulations
- Format tabulation results

Do you know the Measures of Central Tendency?

This package follows the principles of the "tidyverse" and in particular works well with the %>% pipe function. janitor package was built with beginning-to-intermediate R users in mind and is optimized for user-friendliness.

How to clean the datasets in R?

Load library

```
#install.packages("janitor")
library(janitor)
library(dplyr)
```

Getting data

```
data<-read.csv("D:/RStudio/Website/FinData.csv",1)
```

1. Clean column names

First, see the current column names

"First.Name" "Last.Name" "Employee.Status" "Subject" "Hire.Date" "X..Allocated" "Full.time." "do.not.edit……" "Certification" "Certification.1" "Active." "X"

You can use clean_names function for cleaning the data set column names.

```
clean<-clean_names(data)
colnames(clean)
```

"first_name" "last_name" "employee_status" "subject" "hire_date" "x_allocated" "full_time" "do_not_edit" "certification" "certification_1" "active" "x"

How to measure Quality Control of the product?

2. tabyl function

tabyl function is used for easy tabulations (frequency tables and crosstabs)

```
tabyl(clean,employee_status)
employee_status   n    percent
                  5   0.29411765
Administration    1   0.05882353
Coach             2   0.11764706
Teacher           9   0.52941176
```

3. Adorn function

Adorn function is used for formatting the output.

```
clean %>% tabyl(employee_status) %>% adorn_pct_formatting(digits
=2,affix_sign=TRUE)
employee_status    n    percent
                   5     29.41%
  Administration   1     5.88%
        Coach      2     11.76%
        acher      9     52.94%
clean %>% tabyl(employee_status, full_time) %>% adorn_totals()
employee_status   No Yes emptystring_
                  0   0            5
  Administration  0   1            0
           Coach  2   0            0
         Teacher  3   6            0
           Total  5   7            5
clean %>% tabyl(employee_status, full_time) %>% adorn_totals(where =
"col")
employee_status No Yes emptystring_ Total
                0   0            5     5
Administration  0   1            0     1
         Coach  2   0            0     2
       Teacher  3   6            0     9
clean %>% tabyl(employee_status, full_time) %>% adorn_totals(where =
c("row","col"))
employee_status No Yes emptystring_ Total
                  0   0            5     5
  Administration  0   1            0     1
           Coach  2   0            0     2
         Teacher  3   6            0     9
           Total  5   7            5    17
clean %>% tabyl(employee_status, full_time) %>%
adorn_totals("row") %>%
adorn_percentages("row") %>%
adorn_pct_formatting() %>%
adorn_ns()
employee_status       No        Yes    emptystring_
                   0.0% (0)   0.0% (0)   100.0% (5)
  Administration   0.0% (0)  100.0% (1)    0.0% (0)
           Coach 100.0% (2)  0.0% (0)     0.0% (0)
         Teacher  33.3% (3)   66.7% (6)    0.0% (0)
```

```
          Total   29.4% (5)  41.2% (7)    29.4% (5)
```

When you use adorn_ns("front") count column will display as first.

```
clean %>% tabyl(employee_status, full_time) %>%
adorn_totals("row") %>%
adorn_percentages("row") %>%
adorn_pct_formatting() %>%
adorn_ns("front")
 employee_status       No         Yes    emptystring_
                  0   (0.0%) 0   (0.0%)   5 (100.0%)
  Administration 0   (0.0%) 1 (100.0%)   0   (0.0%)
           Coach 2 (100.0%) 0   (0.0%)   0   (0.0%)
         Teacher 3  (33.3%) 6  (66.7%)   0   (0.0%)
           Total 5  (29.4%) 7  (41.2%)   5  (29.4%)
clean %>% tabyl(employee_status,full_time,subject)
```

How to do data reshape in R?

```
employee_status No Yes emptystring_
                   0   0            5
   Administration  0   0            0
            Coach  1   0            0
          Teacher  0   0            0
 $#REF!
  employee_status No Yes emptystring_
                   0   0            0
   Administration  0   0            0
            Coach  0   0            0
          Teacher  0   1            0
 $Basketball
  employee_status No Yes emptystring_
                   0   0            0
   Administration  0   0            0
            Coach  1   0            0
          Teacher  0   0            0
```

4. Remove empty column or rows

Suppose if you want to remove the column or row if contain completely empty, then you can use remove_empty function.

```
clean_x<-clean %>% remove_empty(whic=c("rows"))
clean_x<-clean %>% remove_empty(whic=c("cols"))
```

5. Remove duplicate records

If you want remove duplicate records, then get_dupes will come handy.

```
clean %>% get_dupes(first_name)
clean %>% get_dupes(first_name,certification)
first_name certification dupe_count last_name employee_status   subject
1                                   5
2                                   5
```

```
3                                    5
4                                    5
5                                    5
6 Chien-Shiung  Science 6-12         2        Wu         Teacher
Physics
7 Chien-Shiung  Science 6-12         2        Wu         Teacher
Chemistry
8        Jason   Physical ed         2    Bourne
Teacher        PE
9        Jason   Physical ed         2    Bourne         Teacher
Drafting
hire_date x_allocated full_time do_not_edit certification_1 active  x
1                                        NA                       NA
2                                        NA                       NA
3                                        NA                       NA
4                                        NA                       NA
5                                        NA                       NA
6     11037         50%       Yes          NA         Physics    YES NA
7     11037         50%       Yes          NA         Physics    YES NA
8     39690         75%       Yes          NA         Theater    YES NA
9 1/14/2019         25%       Yes          NA         Theater    YES NA
```

6. Date Format Numeric to Date

Most probably you are experience date issues in r when you are loading from the excel file date column will automatically convert into a numeric form or in excel itself it's displayed as numerical values. Based on excel_numeric_to_date you can easily resolve these issues.

```
excel_numeric_to_date(41103)
"2012-07-13"
```