

Data-607 Week-6 Assignment

Student Name : Sachid Deshmukh

Date : 10/04/2018

- GitHub Location for rmd file
 - GitHub Location for pdf file
 - RPub's location of published file
 - CSV file-1
 - CSV file-2
 - CSV file-3
-

1] Library Initialization

```
library(tidyr)
```

```
## Warning: package 'tidyr' was built under R version 3.4.3
```

```
library(dplyr)
```

```
## Warning: package 'dplyr' was built under R version 3.4.3
```

```
##
```

```
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
```

```
##
```

```
## filter, lag
```

```
## The following objects are masked from 'package:base':
```

```
##
```

```
## intersect, setdiff, setequal, union
```

```
library(ggplot2)
```

2] Data Set-1 Analysis

Read Crime Rate data for three countries reported in year 1999

```
Crime.Rate = read.csv("https://raw.githubusercontent.com/mlforsachid/MSDSQ1/master/Data607/Week6/CrimeR")
```

Preview Data

```
head(Crime.Rate)
```

```
##      country year  crimeinfo      value
## 1 Afghanistan 1999      cases        745
## 2 Afghanistan 1999 population 19987071
## 3      Brazil 1999      cases    37737
## 4      Brazil 1999 population 172006362
## 5      China 1999      cases    212258
## 6      China 1999 population 1272915272
```

```
str(Crime.Rate)
```

```
## 'data.frame':    6 obs. of  4 variables:
## $ country : chr  "Afghanistan" "Afghanistan" "Brazil" "Brazil" ...
## $ year : int  1999 1999 1999 1999 1999 1999
## $ crimeinfo: chr  "cases" "population" "cases" "population" ...
## $ value : int  745 19987071 37737 172006362 212258 1272915272
```

Note that crimeinfo column is stacked. Cases indicates the crime cases reported for a specific country and population indicates total population of the country. Let's spread crimeinfo column

```
Crime.Rate = tidyr::spread(Crime.Rate, crimeinfo, value)
```

Preview unpivoted Crime Rate

```
head(Crime.Rate)
```

```
##      country year  cases population
## 1 Afghanistan 1999    745    19987071
## 2      Brazil 1999  37737    172006362
## 3      China 1999 212258    1272915272
```

Note how crimerate column is spread based on categories. This also flattened the whole data frame.

3] Data Set-2 Analysis

Read Student Grade data for three subjects

```
Stu.Grade = read.csv("https://raw.githubusercontent.com/mlforsachid/MSDSQ1/master/Data607/Week6/Student")
```

Preview Data

```
head(Stu.Grade)
```

```
##   name math science history
## 1 James  68      56      80
## 2   Bob  90      50      67
## 3  Amit  45      89      90
```

```
str(Stu.Grade)
```

```
## 'data.frame':   3 obs. of  4 variables:
## $ name      : chr  "James" "Bob" "Amit"
## $ math      : int   68 90 45
## $ science: int   56 50 89
## $ history: int   80 67 90
```

Note that each grades for a particular subject are on different column. Let's create a single Subject column.

```
Stu.Grade = tidyr::gather(Stu.Grade, "subject", "grades", 2:4)
```

Preview unpivoted Crime Rate

```
head(Stu.Grade)
```

```
##   name subject grades
## 1 James    math     68
## 2  Bob     math     90
## 3  Amit     math     45
## 4 James science    56
## 5  Bob     science   50
## 6  Amit     science   89
```

Note how individual columns for subject are collapsed into single column. The values are captured under newly created grades column

4] Data Set-3 Analysis

Read City Temperature data for three cities

```
City.Temp = read.csv("https://raw.githubusercontent.com/mlforsachid/MSDSQ1/master/Data607/Week6/CityTemp.csv")
```

Preview Data

```
head(City.Temp)
```

```
##      city      date temp
## 1 Redmond 10/01/2018   40
## 2 Bellevue 10/02/2018   38
## 3 Seattle 10/03/2018   42
```

```
str(City.Temp)
```

```
## 'data.frame':   3 obs. of  3 variables:
## $ city: chr  "Redmond" "Bellevue" "Seattle"
## $ date: chr  "10/01/2018" "10/02/2018" "10/03/2018"
## $ temp: int   40 38 42
```

Note Date column. Let's separate Month, Day and Year into separate columns

```
City.Temp = tidyr::separate(City.Temp, "date", c("month", "day", "year"), sep="/")
```

Preview unpivoted Crime Rate

```
head(City.Temp)
```

```
##           city month day year temp
## 1  Redmond      10   01 2018   40
## 2  Bellevue      10   02 2018   38
## 3   Seattle      10   03 2018   42
```

Note how date column is splitted across three separate columns (month, day and year)