Assignment 4 - answer

```
my.cereal=read.delim('cereals.CSV',sep=',')
```

Primary Solution

I decided to populate the missing values for the sugar variable

```
#summary(my.cereal)

##will look into sugar
my.cereal[is.na(my.cereal$Sugars),]
```

```
##               Name Manuf Type Calories Protein Fat Sodium Fiber Carbo
## 58 Quaker_Oatmeal      Q    H      100       5   2      0   2.7    NA
##    Sugars Potass Vitamins Shelf Weight Cups   Rating Cold Nabisco Quaker
## 58     NA    110        0     1      1 0.67 50.82839    0       0      1
##    Kelloggs GeneralMills Ralston AHFP
## 58        0            0       0    0
```

I will be populating the missing value by the sugar avarage of the cereals that are of the same type as the cereal that has the missing sugar value. That is, I am choosing ‘Type’ as the categorical variable.

```
#I will populate it by average of the cereals that have the same type
my.type=my.cereal[is.na(my.cereal$Sugars),]$Type


## All the entries where type is the same as the row with the missing item
same.type.cereals=my.cereal[my.cereal$Type==my.type,]

type.average=mean(same.type.cereals$Sugars, na.rm = TRUE)

my.cereal[is.na(my.cereal$Sugars),]$Sugars=type.average
```

Alternative Solution

```
require(magrittr)
```

```
## Loading required package: magrittr
```

```
require(dplyr)
```

```
## Loading required package: dplyr
```

```
## 
## 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
```

```
## To do it for all missing items of various types
my.cereal=read.delim('cereals.CSV',sep=',')

missing_rows=which(is.na(my.cereal$Sugars))

#Showing missing row
my.cereal[missing_rows,]
```

```
##               Name Manuf Type Calories Protein Fat Sodium Fiber Carbo
## 58 Quaker_Oatmeal     Q    H       100       5   2      0   2.7    NA
##    Sugars Potass Vitamins Shelf Weight Cups   Rating Cold Nabisco Quaker
## 58     NA    110        0     1      1 0.67 50.82839    0       0      1
##    Kelloggs GeneralMills Ralston AHFP
## 58        0            0       0    0
```

```
#Fixing missing rows
my.cereal%<>%group_by(Type)%>%mutate(type.average=mean(Sugars, na.rm=TRUE))%>%ungroup()%>%
    mutate(Sugars=ifelse(is.na(Sugars),type.average, Sugars))

#Re-checking missing rows
my.cereal[missing_rows,]
```

```
## # A tibble: 1 x 24
##              Name  Manuf   Type Calories Protein   Fat Sodium Fiber Carbo
##            <fctr> <fctr> <fctr>    <int>   <int> <int>  <int> <dbl> <dbl>
## 1 Quaker_Oatmeal      Q      H       100       5     2      0   2.7    NA
## # ... with 15 more variables: Sugars <dbl>, Potass <int>, Vitamins <int>,
## #   Shelf <int>, Weight <dbl>, Cups <dbl>, Rating <dbl>, Cold <int>,
## #   Nabisco <int>, Quaker <int>, Kelloggs <int>, GeneralMills <int>,
## #   Ralston <int>, AHFP <int>, type.average <dbl>
```