Assignment 4 - CT5102

Data Cleaning and Simple Imputation with Data Frames

The goal of this assignment is to (1) take a copy of the **ggplot2::mpg** data set; (2) insert invalid values (column **cty**) into 10 random records; (3) convert the invalid values to the type NA; (4) calculate the mean **cty** value for each car classl and (5) replace the NA values with the mean in the data set.

The first step is to set the see to 100, and create a copy of **ggplot2::mpg**, and select 10 random observations, using the function **sample()**, and set the **cty** variable to be -999 for each observation.

```
##  [1] 202 102 112 217 206 151 214 198   4  55
```

```
##     manufacturer             model displ year cyl      trans drv  cty hwy
## 202       toyota toyota tacoma 4wd   2.7 1999   4   auto(l4)   4 -999  20
## 102        honda             civic   1.6 1999   4 manual(m5)   f -999  32
## 112      hyundai            sonata   2.4 2008   4 manual(m5)   f -999  31
## 217   volkswagen             jetta   2.0 2008   4 manual(m6)   f -999  29
## 206       toyota toyota tacoma 4wd   4.0 2008   6 manual(m6)   4 -999  18
## 151       nissan    pathfinder 4wd   3.3 1999   6   auto(l4)   4 -999  17
## 214   volkswagen             jetta   2.0 1999   4 manual(m5)   f -999  29
## 198       toyota           corolla   1.8 2008   4   auto(l4)   f -999  35
## 4           audi                a4   2.0 2008   4   auto(av)   f -999  30
## 55         dodge dakota pickup 4wd   4.7 2008   8   auto(l5)   4 -999  12
##     fl      class
## 202  r     pickup
## 102  r subcompact
## 112  r    midsize
## 217  p    compact
## 206  r     pickup
## 151  r        suv
## 214  r    compact
## 198  r    compact
## 4    p    compact
## 55   e     pickup
```

Next, using **lapply** (and converting the result to a data frame), replace all negative values with the symbol NA (note **trans** column omitted for formatting purposes)

```
##     manufacturer             model displ year cyl drv cty hwy fl
## 202       toyota toyota tacoma 4wd   2.7 1999   4   4  NA  20  r
## 102        honda             civic   1.6 1999   4   f  NA  32  r
## 112      hyundai            sonata   2.4 2008   4   f  NA  31  r
## 217   volkswagen             jetta   2.0 2008   4   f  NA  29  p
## 206       toyota toyota tacoma 4wd   4.0 2008   6   4  NA  18  r
## 151       nissan    pathfinder 4wd   3.3 1999   6   4  NA  17  r
## 214   volkswagen             jetta   2.0 1999   4   f  NA  29  r
## 198       toyota           corolla   1.8 2008   4   f  NA  35  r
## 4           audi                a4   2.0 2008   4   f  NA  30  p
## 55         dodge dakota pickup 4wd   4.7 2008   8   4  NA  12  e
```

Next, calculate the mean **cty** for each of the different classes of car (excluding NA values)

```
##    compact    midsize        suv    2seater    minivan     pickup
##   19.93023   18.70000   13.49180   15.40000   15.81818   12.96667
## subcompact
##   20.23529
```

Then, for each row with a missing value, replace the NA with the mean of the car **class**

```
##     manufacturer             model displ year cyl      cty fl      class
## 202       toyota toyota tacoma 4wd   2.7 1999   4 12.96667  r     pickup
## 102        honda             civic   1.6 1999   4 20.23529  r subcompact
## 112      hyundai            sonata   2.4 2008   4 18.70000  r    midsize
## 217   volkswagen             jetta   2.0 2008   4 19.93023  p    compact
## 206       toyota toyota tacoma 4wd   4.0 2008   6 12.96667  r     pickup
## 151       nissan    pathfinder 4wd   3.3 1999   6 13.49180  r        suv
## 214   volkswagen             jetta   2.0 1999   4 19.93023  r    compact
## 198       toyota           corolla   1.8 2008   4 19.93023  r    compact
## 4           audi                a4   2.0 2008   4 19.93023  p    compact
## 55         dodge dakota pickup 4wd   4.7 2008   8 12.96667  e     pickup
```

Finally, confirm that all values are now valid in the data frame by summarising the data.

```
##      manufacturer                model         displ            year
##  dodge     :37    caravan 2wd        : 11   Min.   :1.600   Min.   :1999
##  toyota    :34    ram 1500 pickup 4wd: 10   1st Qu.:2.400   1st Qu.:1999
##  volkswagen:27    civic              :  9   Median :3.300   Median :2004
##  ford      :25    dakota pickup 4wd  :  9   Mean   :3.472   Mean   :2004
##  chevrolet :19    jetta              :  9   3rd Qu.:4.600   3rd Qu.:2008
##  audi      :18    mustang            :  9   Max.   :7.000   Max.   :2008
##  (Other)   :74    (Other)            :177
##       cyl              trans    drv          cty             hwy
##  Min.   :4.000   auto(l4)  :83   4:103   Min.   : 9.00   Min.   :12.00
##  1st Qu.:4.000   manual(m5):58   f:106   1st Qu.:14.00   1st Qu.:18.00
##  Median :6.000   auto(l5)  :39   r: 25   Median :17.00   Median :24.00
##  Mean   :5.889   manual(m6):19           Mean   :16.78   Mean   :23.44
##  3rd Qu.:8.000   auto(s6)  :16           3rd Qu.:19.00   3rd Qu.:27.00
##  Max.   :8.000   auto(l6)  : 6           Max.   :35.00   Max.   :44.00
##                  (Other)   :13
##  fl             class
##  c:  1   2seater   : 5
##  d:  5   compact   :47
##  e:  8   midsize   :41
##  p: 52   minivan   :11
##  r:168   pickup    :33
##          subcompact:35
##          suv       :62
```

Show that the two mean values now differ:

```
mean(mpg$cty)
```

```
## [1] 16.85897
```

```
mean(new_df$cty)
```

```
## [1] 16.78226
```