Project

SeanJ- Volcaetus

4/3/2022

Data

From TidyTuesday URL:https://github.com/rfordatascience/tidytuesday/tree/master/data/2020/2020-07-07

```
coffee_ratings <- readr::read_csv('https://raw.githubusercontent.com/rfordatascience/tidytuesday/master/
```

```
##
## -- Column specification --------------------------------------------------------
## cols(
##   .default = col_character(),
##   total_cup_points = col_double(),
##   number_of_bags = col_double(),
##   aroma = col_double(),
##   flavor = col_double(),
##   aftertaste = col_double(),
##   acidity = col_double(),
##   body = col_double(),
##   balance = col_double(),
##   uniformity = col_double(),
##   clean_cup = col_double(),
##   sweetness = col_double(),
##   cupper_points = col_double(),
##   moisture = col_double(),
##   category_one_defects = col_double(),
##   quakers = col_double(),
##   category_two_defects = col_double(),
##   altitude_low_meters = col_double(),
##   altitude_high_meters = col_double(),
##   altitude_mean_meters = col_double()
## )
## i Use `spec()` for the full column specifications.
```

Quick overview

```
summary(coffee_ratings)
```

```
##  total_cup_points   species             owner           country_of_origin
##  Min.   : 0.00    Length:1339        Length:1339        Length:1339
##  1st Qu.:81.08    Class :character   Class :character   Class :character
##  Median :82.50    Mode  :character   Mode  :character   Mode  :character
##  Mean   :82.09
##  3rd Qu.:83.67
```

```
##  Max.   :90.58
##
##   farm_name          lot_number            mill            ico_number
##  Length:1339        Length:1339        Length:1339        Length:1339
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##
##    company            altitude           region            producer
##  Length:1339        Length:1339        Length:1339        Length:1339
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##
##  number_of_bags    bag_weight        in_country_partner harvest_year
##  Min.   :   0.0   Length:1339        Length:1339        Length:1339
##  1st Qu.:  14.0   Class :character   Class :character   Class :character
##  Median : 175.0   Mode  :character   Mode  :character   Mode  :character
##  Mean   : 154.2
##  3rd Qu.: 275.0
##  Max.   :1062.0
##
##  grading_date         owner_1            variety          processing_method
##  Length:1339        Length:1339        Length:1339        Length:1339
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##
##      aroma            flavor       aftertaste       acidity            body
##  Min.   :0.000   Min.   :0.00   Min.   :0.000   Min.   :0.000   Min.   :0.000
##  1st Qu.:7.420   1st Qu.:7.33   1st Qu.:7.250   1st Qu.:7.330   1st Qu.:7.330
##  Median :7.580   Median :7.58   Median :7.420   Median :7.580   Median :7.500
##  Mean   :7.567   Mean   :7.52   Mean   :7.401   Mean   :7.536   Mean   :7.517
##  3rd Qu.:7.750   3rd Qu.:7.75   3rd Qu.:7.580   3rd Qu.:7.750   3rd Qu.:7.670
##  Max.   :8.750   Max.   :8.83   Max.   :8.670   Max.   :8.750   Max.   :8.580
##
##     balance        uniformity       clean_cup        sweetness
##  Min.   :0.000   Min.   : 0.000   Min.   : 0.000   Min.   : 0.000
##  1st Qu.:7.330   1st Qu.:10.000   1st Qu.:10.000   1st Qu.:10.000
##  Median :7.500   Median :10.000   Median :10.000   Median :10.000
##  Mean   :7.518   Mean   : 9.835   Mean   : 9.835   Mean   : 9.857
##  3rd Qu.:7.750   3rd Qu.:10.000   3rd Qu.:10.000   3rd Qu.:10.000
##  Max.   :8.750   Max.   :10.000   Max.   :10.000   Max.   :10.000
##
##  cupper_points       moisture       category_one_defects    quakers
##  Min.   : 0.000   Min.   :0.00000   Min.   : 0.0000      Min.   : 0.0000
##  1st Qu.: 7.250   1st Qu.:0.09000   1st Qu.: 0.0000      1st Qu.: 0.0000
##  Median : 7.500   Median :0.11000   Median : 0.0000      Median : 0.0000
```

```
##  Mean   : 7.503   Mean   :0.08838   Mean   : 0.4795      Mean   : 0.1734
##  3rd Qu.: 7.750   3rd Qu.:0.12000   3rd Qu.: 0.0000      3rd Qu.: 0.0000
##  Max.   :10.000   Max.   :0.28000   Max.   :63.0000      Max.   :11.0000
##                                                          NA's   :1
##     color           category_two_defects  expiration        certification_body
##  Length:1339        Min.   : 0.000       Length:1339        Length:1339
##  Class :character   1st Qu.: 0.000       Class :character   Class :character
##  Mode  :character   Median : 2.000       Mode  :character   Mode  :character
##                     Mean   : 3.556
##                     3rd Qu.: 4.000
##                     Max.   :55.000
##
##  certification_address certification_contact unit_of_measurement
##  Length:1339           Length:1339           Length:1339
##  Class :character      Class :character      Class :character
##  Mode  :character      Mode  :character      Mode  :character
##
##
##
##
##  altitude_low_meters altitude_high_meters altitude_mean_meters
##  Min.   :     1      Min.   :     1       Min.   :     1
##  1st Qu.:  1100      1st Qu.:  1100       1st Qu.:  1100
##  Median :  1311      Median :  1350       Median :  1311
##  Mean   :  1751      Mean   :  1799       Mean   :  1775
##  3rd Qu.:  1600      3rd Qu.:  1650       3rd Qu.:  1600
##  Max.   :190164      Max.   :190164       Max.   :190164
##  NA's   :230         NA's   :230          NA's   :230
```

A few NA's.

1 within quakers, and 230 in Altitude low/high/mean

Check what is happening in the rest of the data set

Count of NA's per coloumn

```
apply(X=is.na(coffee_ratings), MARGIN = 2, FUN = sum)
```

```
##      total_cup_points               species                 owner
##                     0                     0                     7
##     country_of_origin             farm_name           lot_number
##                     1                   359                  1063
##                  mill             ico_number               company
##                   315                   151                   209
##              altitude                region              producer
##                   226                    59                   231
##        number_of_bags           bag_weight    in_country_partner
##                     0                     0                     0
##          harvest_year          grading_date               owner_1
##                    47                     0                     7
##               variety     processing_method                 aroma
##                   226                   170                     0
##                flavor            aftertaste               acidity
##                     0                     0                     0
```

```
##                   body                balance            uniformity
##                      0                      0                     0
##             clean_cup              sweetness         cupper_points
##                      0                      0                     0
##              moisture   category_one_defects               quakers
##                      0                      0                     1
##                 color   category_two_defects            expiration
##                   218                      0                     0
##    certification_body certification_address certification_contact
##                      0                      0                     0
##   unit_of_measurement   altitude_low_meters  altitude_high_meters
##                      0                    230                   230
##  altitude_mean_meters
##                   230
```

I will be just removing some of the columns with many missing vlaues, for instance farm_name.

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --
```

```
## v ggplot2 3.3.4     v purrr   0.3.4
## v tibble  3.1.2     v dplyr   1.0.7
## v tidyr   1.1.3     v stringr 1.4.0
## v readr   1.4.0     v forcats 0.5.1
```

```
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

Removal of columns

```
coffee = coffee_ratings%>%
  select(-farm_name,-lot_number,-mill,-ico_number,-altitude,
         -altitude_low_meters,-altitude_high_meters,-producer,-company,
         -expiration,-certification_address,-owner_1,-grading_date)
apply(X=is.na(coffee), MARGIN = 2, FUN = sum)
```

```
##      total_cup_points               species                 owner
##                     0                     0                     7
##     country_of_origin                region         number_of_bags
##                     1                    59                     0
##            bag_weight    in_country_partner          harvest_year
##                     0                     0                    47
##               variety     processing_method                 aroma
##                   226                   170                     0
##                flavor            aftertaste               acidity
##                     0                     0                     0
##                  body               balance            uniformity
##                     0                     0                     0
##             clean_cup             sweetness         cupper_points
##                     0                     0                     0
##              moisture   category_one_defects               quakers
##                     0                     0                     1
##                 color   category_two_defects    certification_body
##                   218                     0                     0
## certification_contact   unit_of_measurement   altitude_mean_meters
```

```
##                     0                     0                   230
```

```
view(coffee)
```

```
try = na.omit(coffee)
```

```
#df only na's
NAs=coffee_ratings%>%
  select(everything())%>%
  filter(is.na(quakers | altitude_mean_meters | altitude_high_meters | altitude_low_meters))
#take a look at missing values within rows
NAs
```

```
## # A tibble: 203 x 43
##    total_cup_points species owner  country_of_origin farm_name  lot_number mill
##               <dbl> <chr>   <chr>  <chr>             <chr>      <chr>      <chr>
##  1             88.8 Arabica ji-ae~ Brazil            <NA>       <NA>       <NA>
##  2             88.8 Arabica hugo ~ Peru              <NA>       <NA>       hvc
##  3             87.3 Arabica ethio~ Ethiopia          <NA>       <NA>       <NA>
##  4             87.1 Arabica ji-ae~ Ethiopia          <NA>       <NA>       <NA>
##  5             86.9 Arabica ethio~ Ethiopia          <NA>       <NA>       <NA>
##  6             86.6 Arabica nora ~ Nicaragua         <NA>       <NA>       bene~
##  7             86.5 Arabica speci~ Tanzania, United~ <NA>       <NA>       <NA>
##  8             86.2 Arabica kona ~ United States (H~ <NA>       <NA>       <NA>
##  9             86.2 Arabica kona ~ United States (H~ kona paci~ <NA>       <NA>
## 10             86.2 Arabica ethio~ Ethiopia          phone num~ <NA>       <NA>
## # ... with 193 more rows, and 36 more variables: ico_number <chr>,
## #   company <chr>, altitude <chr>, region <chr>, producer <chr>,
## #   number_of_bags <dbl>, bag_weight <chr>, in_country_partner <chr>,
## #   harvest_year <chr>, grading_date <chr>, owner_1 <chr>, variety <chr>,
## #   processing_method <chr>, aroma <dbl>, flavor <dbl>, aftertaste <dbl>,
## #   acidity <dbl>, body <dbl>, balance <dbl>, uniformity <dbl>,
## #   clean_cup <dbl>, sweetness <dbl>, cupper_points <dbl>, moisture <dbl>,
## #   category_one_defects <dbl>, quakers <dbl>, color <chr>,
## #   category_two_defects <dbl>, expiration <chr>, certification_body <chr>,
## #   certification_address <chr>, certification_contact <chr>,
## #   unit_of_measurement <chr>, altitude_low_meters <dbl>,
## #   altitude_high_meters <dbl>, altitude_mean_meters <dbl>
```

```
#df with no na's
coffee=coffee_ratings%>%
  select(everything())%>%
  filter((!is.na(quakers | altitude_mean_meters | altitude_high_meters | altitude_low_meters)))
```

```
old = count(coffee_ratings)
new = count(coffee)
removed = old - new
pct.Kept = new / old
pct.Removed = removed / old
print(c(old,new,removed,pct.Kept,pct.Removed))
```

```
## $n
## [1] 1339
##
## $n
## [1] 1136
```

```
##
## $n
## [1] 203
##
## $n
## [1] 0.8483943
##
## $n
## [1] 0.1516057
```

I would prefer to have less removed, but this captures at least 80% of the coffee rating data. If the altitude does not show to have an effect on the coffee rating, then the removal of the altitude columns from the original data will be done. Then re-running analyis on this set will occur at that time.

```
coffee
```

```
## # A tibble: 1,136 x 43
##    total_cup_points species owner   country_of_orig~ farm_name  lot_number mill
##               <dbl> <chr>   <chr>   <chr>            <chr>      <chr>      <chr>
##  1             90.6 Arabica metad ~ Ethiopia         "metad pl~ <NA>       meta~
##  2             89.9 Arabica metad ~ Ethiopia         "metad pl~ <NA>       meta~
##  3             89.8 Arabica ground~ Guatemala        "san marc~ <NA>       <NA>
##  4             89   Arabica yidnek~ Ethiopia         "yidnekac~ <NA>       wole~
##  5             88.8 Arabica metad ~ Ethiopia         "metad pl~ <NA>       meta~
##  6             88.7 Arabica ethiop~ Ethiopia         "aolme"    <NA>       c.p.~
##  7             88.4 Arabica ethiop~ Ethiopia         "aolme"    <NA>       c.p.~
##  8             88.2 Arabica diamon~ Ethiopia         "tulla co~ <NA>       tull~
##  9             88.1 Arabica mohamm~ Ethiopia         "fahem co~ <NA>       <NA>
## 10             87.9 Arabica cqi q ~ United States    "el filo"  <NA>       <NA>
## # ... with 1,126 more rows, and 36 more variables: ico_number <chr>,
## #   company <chr>, altitude <chr>, region <chr>, producer <chr>,
## #   number_of_bags <dbl>, bag_weight <chr>, in_country_partner <chr>,
## #   harvest_year <chr>, grading_date <chr>, owner_1 <chr>, variety <chr>,
## #   processing_method <chr>, aroma <dbl>, flavor <dbl>, aftertaste <dbl>,
## #   acidity <dbl>, body <dbl>, balance <dbl>, uniformity <dbl>,
## #   clean_cup <dbl>, sweetness <dbl>, cupper_points <dbl>, moisture <dbl>,
## #   category_one_defects <dbl>, quakers <dbl>, color <chr>,
## #   category_two_defects <dbl>, expiration <chr>, certification_body <chr>,
## #   certification_address <chr>, certification_contact <chr>,
## #   unit_of_measurement <chr>, altitude_low_meters <dbl>,
## #   altitude_high_meters <dbl>, altitude_mean_meters <dbl>
```