Data imputation on Kaggle's Craft Beers Dataset

Daniel Navarro

2025-08-02

Goal

In order to check Multivariate Imputation by Chained Equations we take the Craft Beers Dataset from Kaggle https://www.kaggle.com/datasets/nickhould/craft-cans/data with contains two datasets with 2410 different beers dataset and 558 in the dataset of breweries from the USA.

Datasets are incomplete, lacking 1067 data points distributed in different fields of the beers and 0 of the breweries.

We cannot afford to discard 1067 rows out of a total of 2410 and must find a way to impute this data, but, how is the missing data distributed?

The datasets are described a continuation with abv meaning alcoholic content by volume, ibu is international bittering units, the rest of variables are self explained:

```
summary(original_beers)
```

```
##        X               abv               ibu               id
##  Min.   :   0.0   Min.   :0.00100   Min.   :  4.00   Min.   :   1.0
##  1st Qu.: 602.2   1st Qu.:0.05000   1st Qu.: 21.00   1st Qu.: 808.2
##  Median :1204.5   Median :0.05600   Median : 35.00   Median :1453.5
##  Mean   :1204.5   Mean   :0.05977   Mean   : 42.71   Mean   :1431.1
##  3rd Qu.:1806.8   3rd Qu.:0.06700   3rd Qu.: 64.00   3rd Qu.:2075.8
##  Max.   :2409.0   Max.   :0.12800   Max.   :138.00   Max.   :2692.0
##                   NA's   :62        NA's   :1005
##      name              style             brewery_id        ounces
##  Length:2410        Length:2410        Min.   :  0.0   Min.   : 8.40
##  Class :character   Class :character   1st Qu.: 93.0   1st Qu.:12.00
##  Mode  :character   Mode  :character   Median :205.0   Median :12.00
##                                        Mean   :231.7   Mean   :13.59
##                                        3rd Qu.:366.0   3rd Qu.:16.00
##                                        Max.   :557.0   Max.   :32.00
##
```

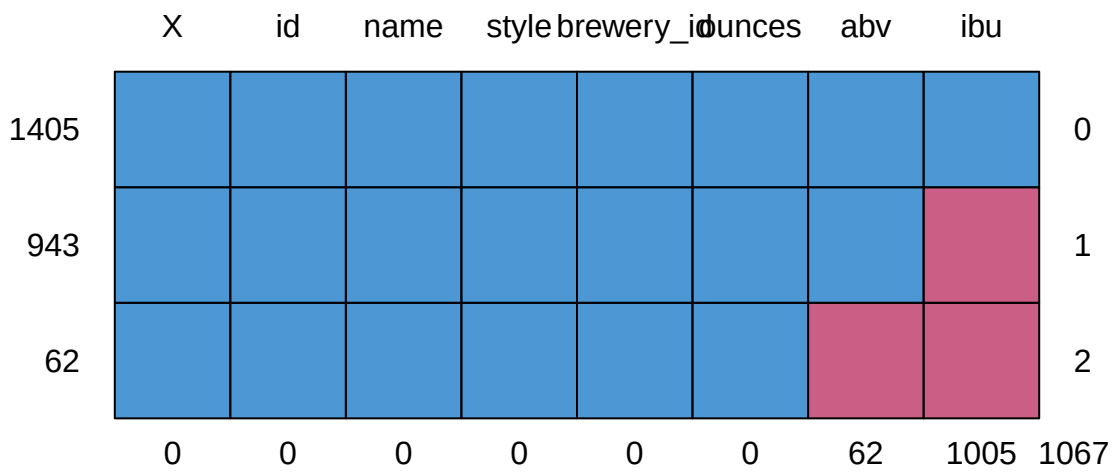
```
summary(original_breweries)
```

```
##        X             name               city              state
##  Min.   :  0.0   Length:558         Length:558         Length:558
##  1st Qu.:139.2   Class :character   Class :character   Class :character
##  Median :278.5   Mode  :character   Mode  :character   Mode  :character
##  Mean   :278.5
##  3rd Qu.:417.8
##  Max.   :557.0
```

Data Imputation

We will check how are missing datapoints distributed between fields in the beers dataset.

```
md.pattern(original_beers)
```




```
##      X id name style brewery_id ounces abv  ibu
## 1405 1  1    1     1          1      1   1    1    0
## 943  1  1    1     1          1      1   1    0    1
## 62   1  1    1     1          1      1   0    0    2
##      0  0    0     0          0      0  62 1005 1067
```

Only abv and ibu variables present missing values, with 943 rows missing only in ibu and 62 rows missing two variables, abv and ibu.

Simple data imputation with the mean

This is the very common practice of filling missing data points with the mean of the existing values.

```
beers_completed_mean <- original_beers
imp_mean <- mice(beers_completed_mean, method = 'mean', m = 1, maxit = 1)
```

```
##
##  iter imp variable
##   1   1  abv  ibu
```

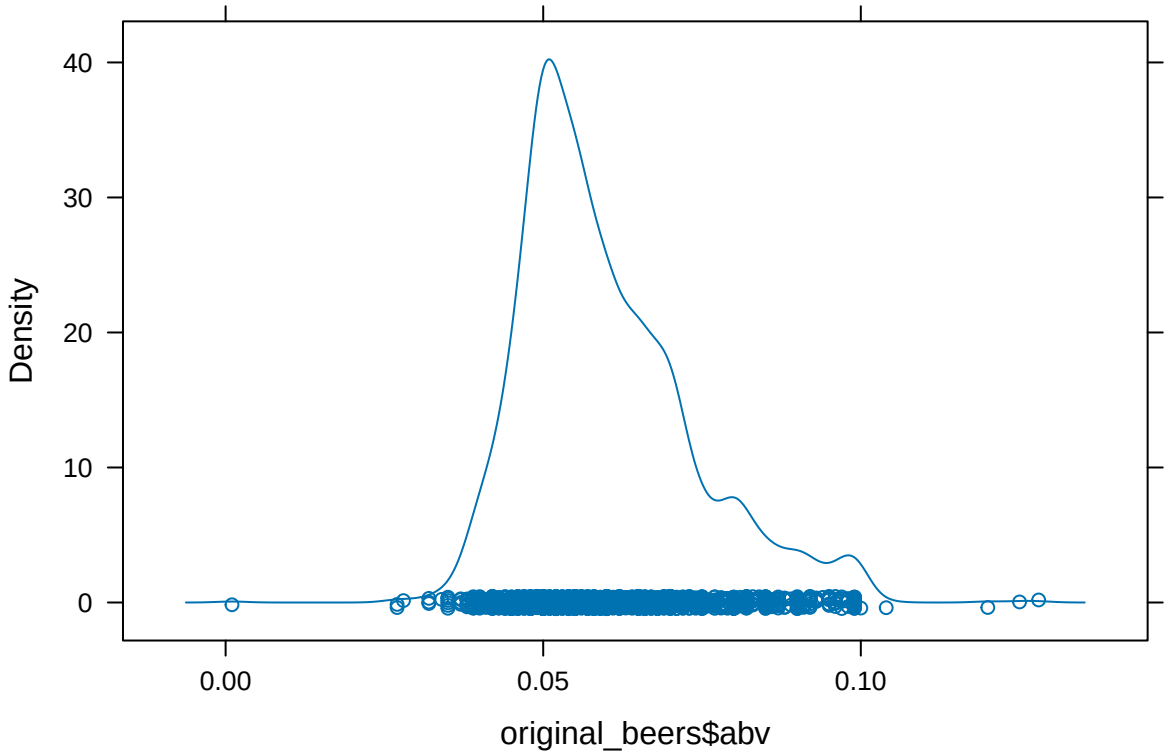
```
beers_completed_mean <- complete(imp_mean)
summary(beers_completed_mean)
```

```
##        X               abv               ibu               id
##  Min.   :   0.0   Min.   :0.00100   Min.   :  4.00   Min.   :   1.0
##  1st Qu.: 602.2   1st Qu.:0.05000   1st Qu.: 30.00   1st Qu.: 808.2
##  Median :1204.5   Median :0.05700   Median : 42.71   Median :1453.5
##  Mean   :1204.5   Mean   :0.05977   Mean   : 42.71   Mean   :1431.1
##  3rd Qu.:1806.8   3rd Qu.:0.06700   3rd Qu.: 42.71   3rd Qu.:2075.8
##  Max.   :2409.0   Max.   :0.12800   Max.   :138.00   Max.   :2692.0
##      name              style             brewery_id        ounces
##  Length:2410        Length:2410        Min.   :  0.0   Min.   : 8.40
##  Class :character   Class :character   1st Qu.: 93.0   1st Qu.:12.00
##  Mode  :character   Mode  :character   Median :205.0   Median :12.00
##                                        Mean   :231.7   Mean   :13.59
##                                        3rd Qu.:366.0   3rd Qu.:16.00
##                                        Max.   :557.0   Max.   :32.00
```

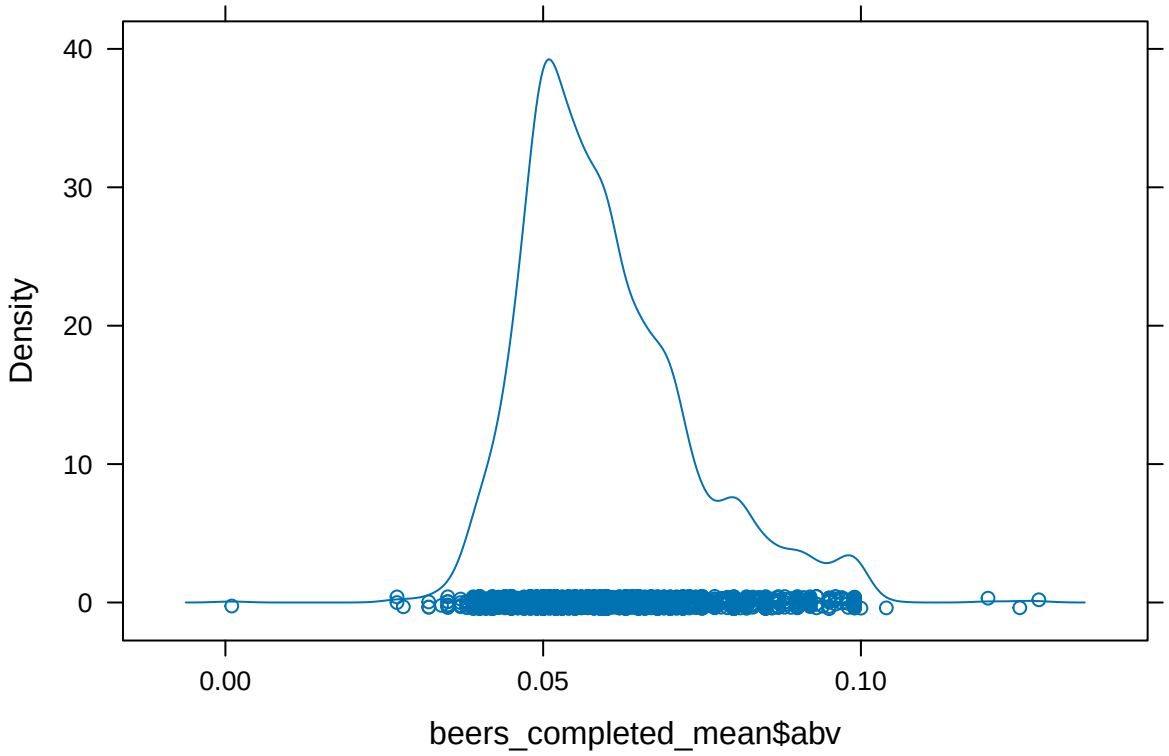
After a first imputation of the missing data points, using only average we have completed all the missing data in all variables.

After comparing the densities of both variables, pre and post imputations using the mean, abv is still consistent with the original set, however ibu has been over centered. This is the effect of using the average on almost half of the data for ibu variable, so the mean is not a balanced way to input this data.

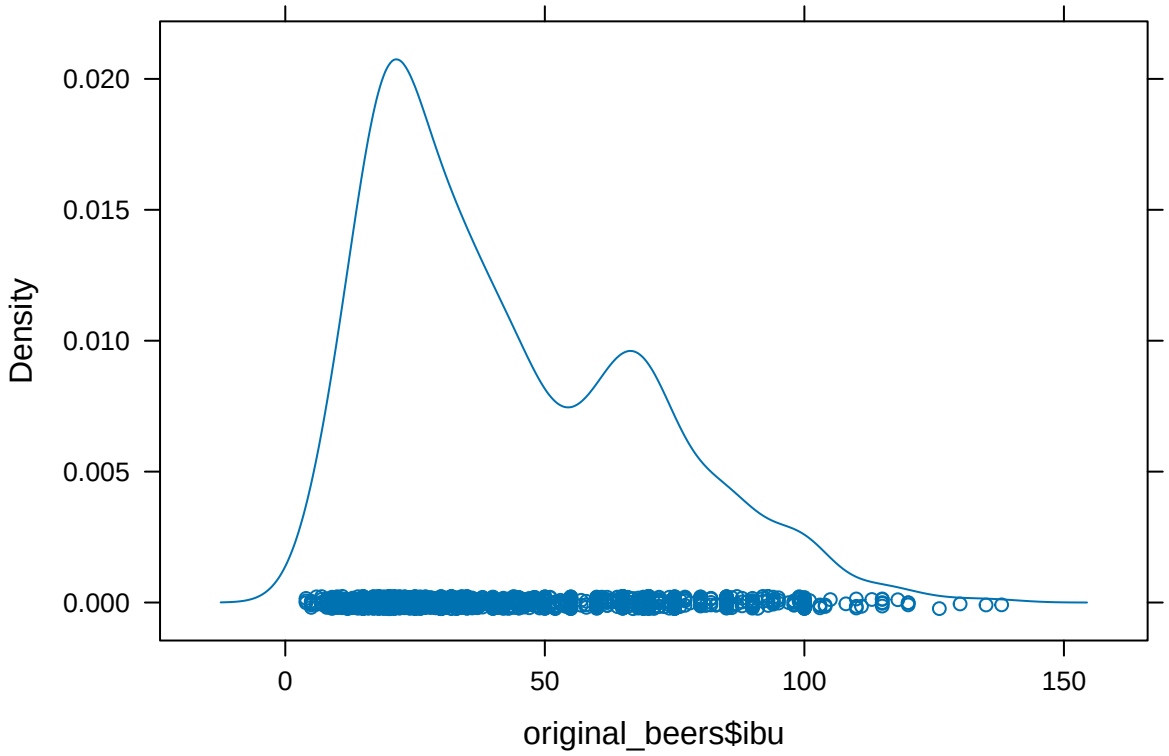
```
densityplot(original_beers$abv)
```



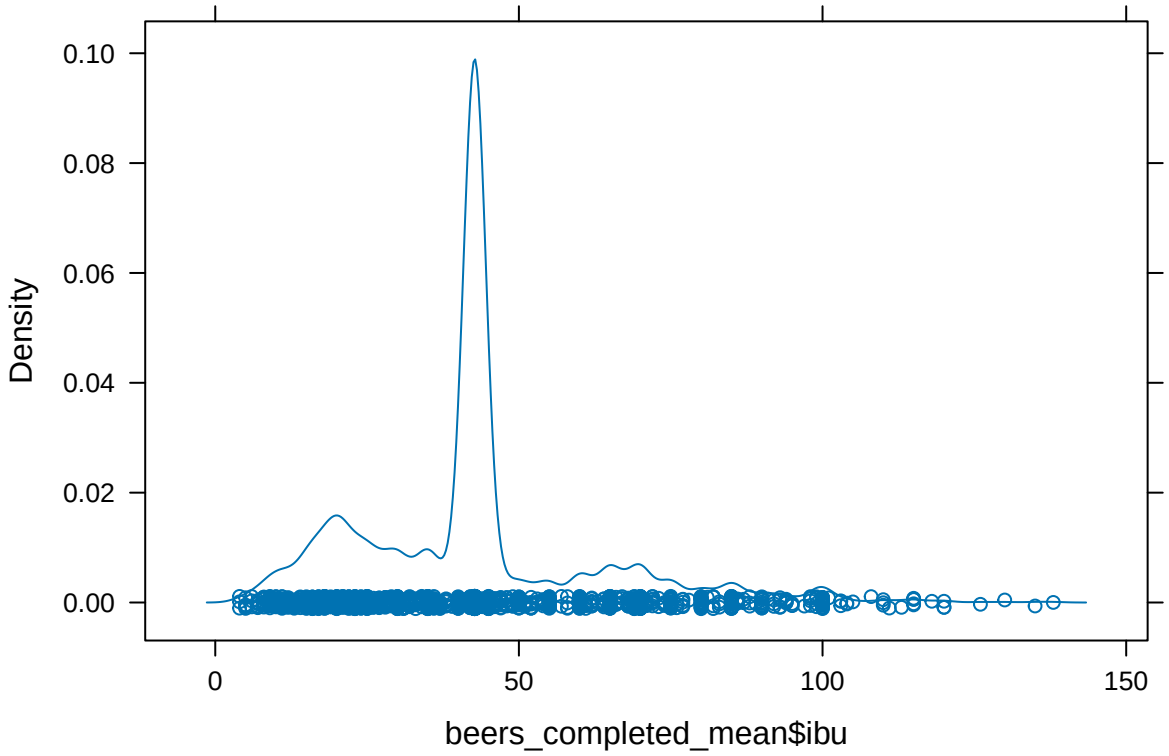

```
densityplot(beers_completed_mean$abv)
```




```
densityplot(original_beers$ibu)
```



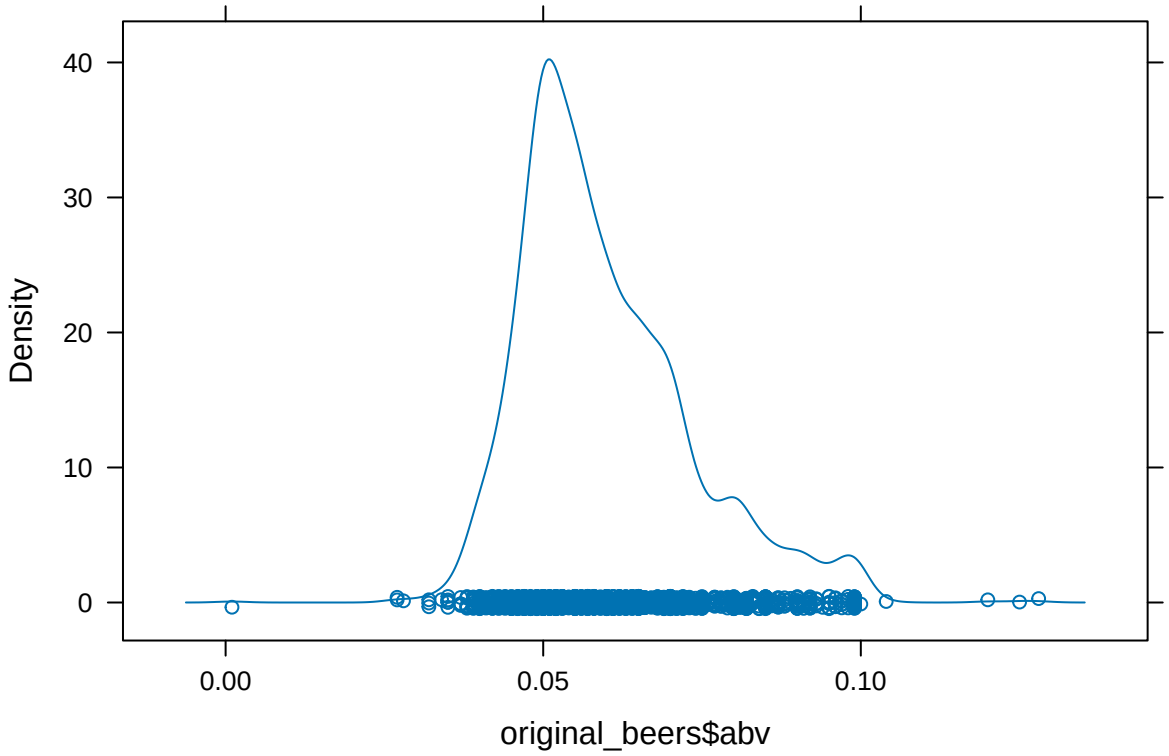

```
densityplot(beers_completed_mean$ibu)
```



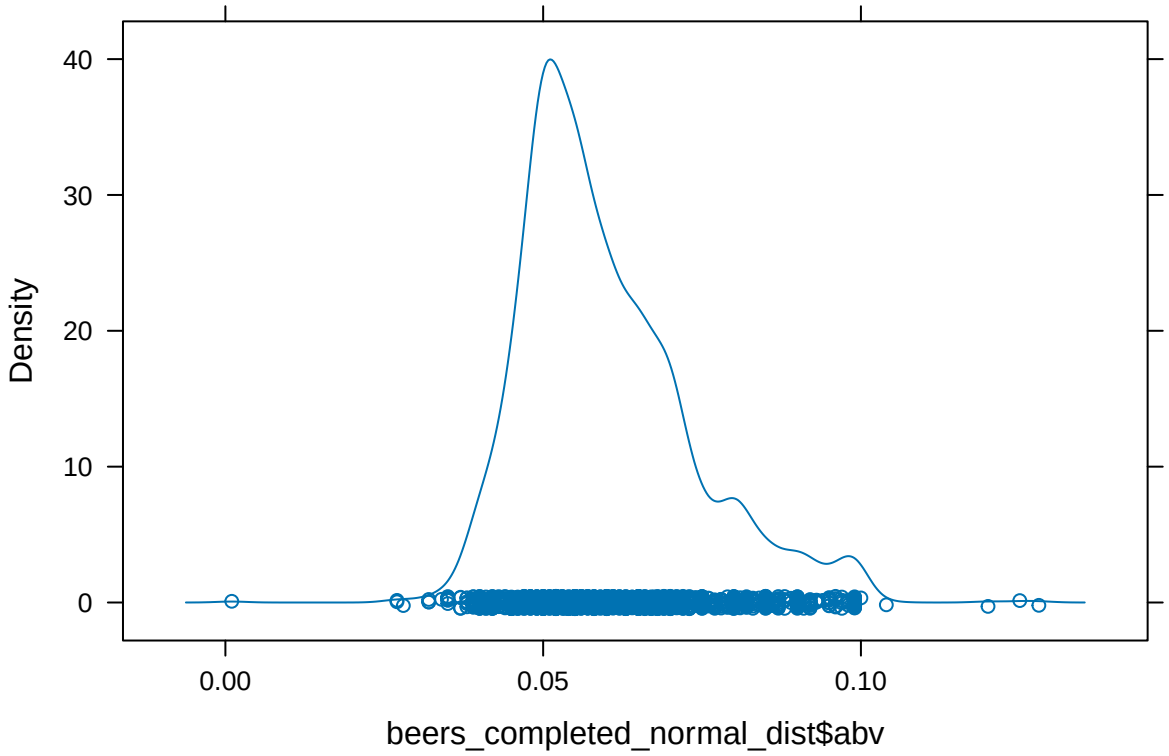
Will test another method of imputation:

Data imputation with normal distribution

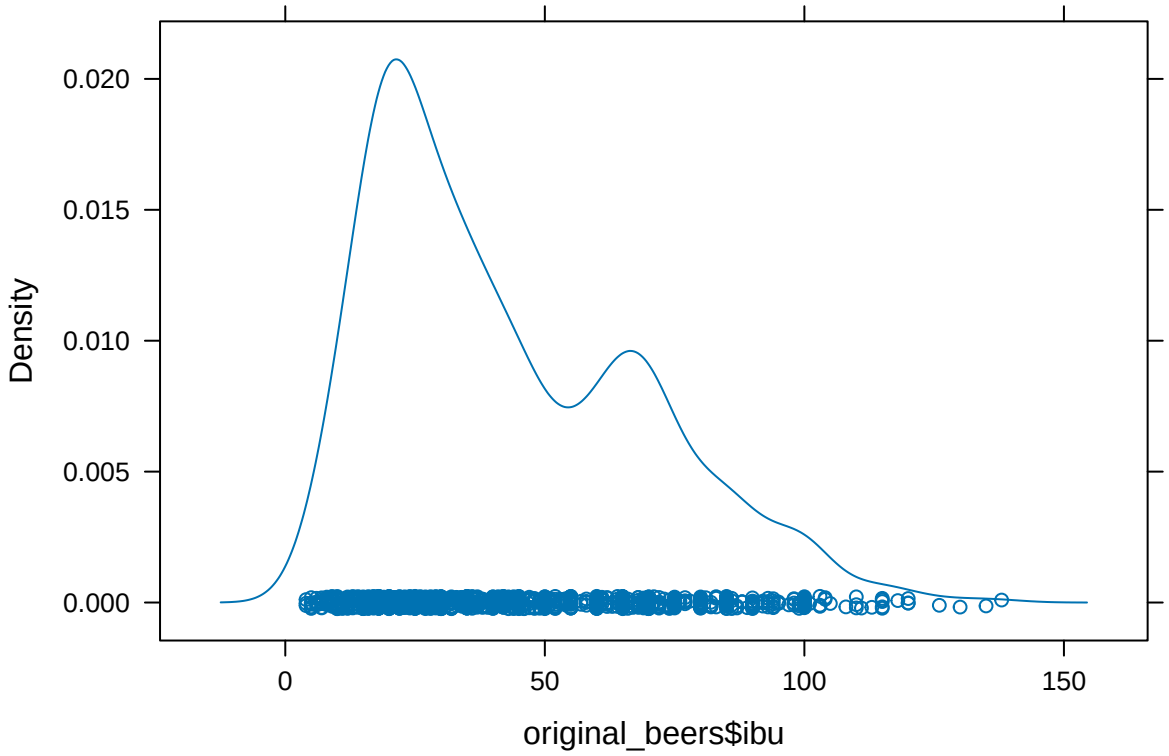
```
par(mfrow = c(2, 2))
densityplot(original_beers$abv)
```



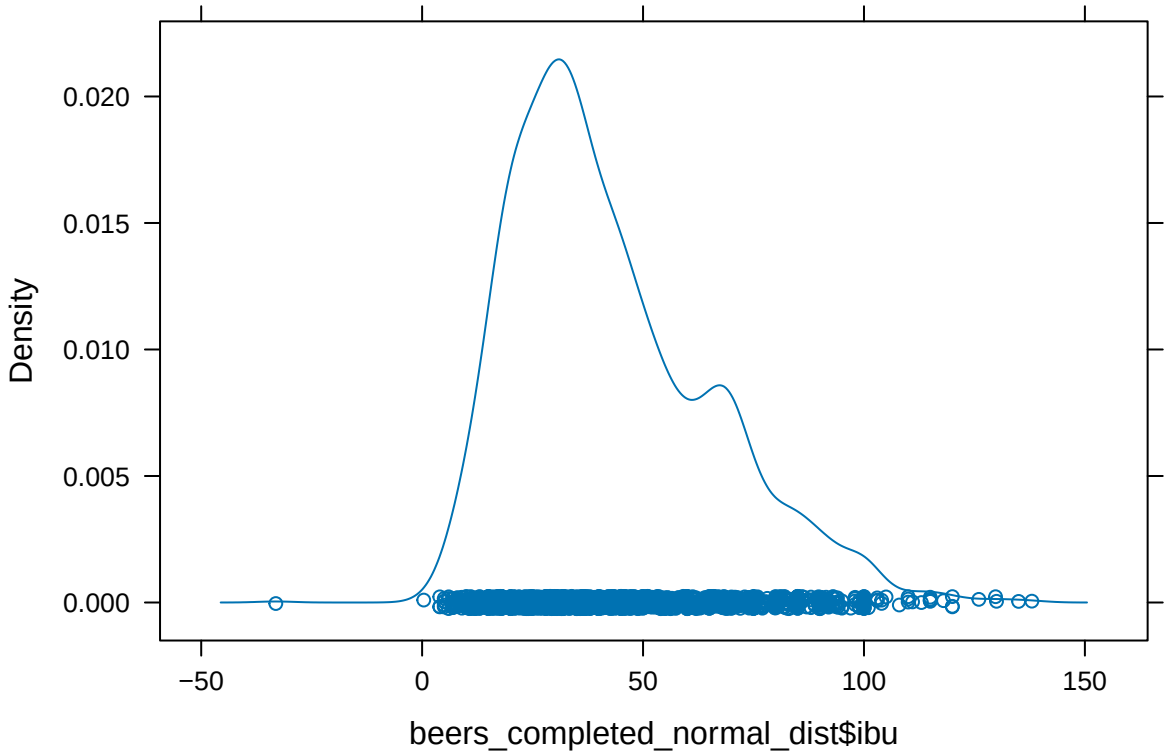

```
densityplot(beers_completed_normal_dist$abv)
```




```
densityplot(original_beers$ibu)
```




```
densityplot(beers_completed_normal_dist$ibu)
```

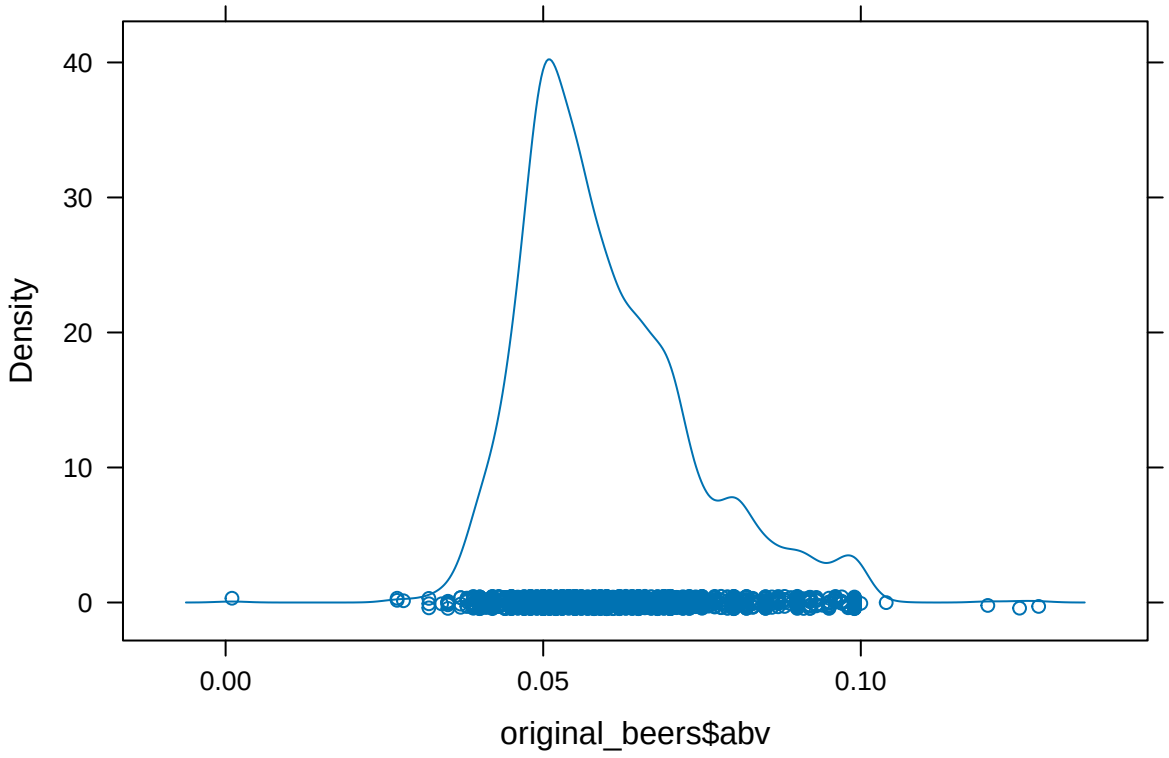


We obtained a completed data set with a distribution coherent with the original data.

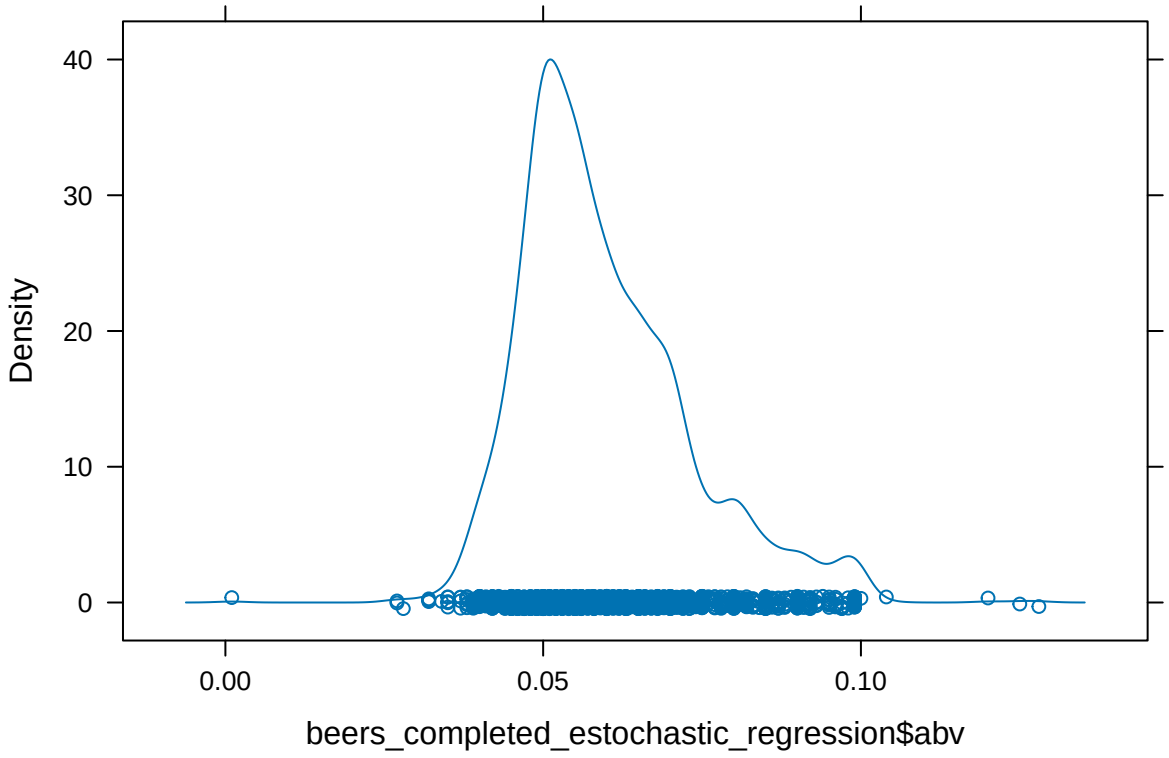
Although the data is now satisfactory, for the sake of practice and curiosity we can try other method, this time:

Data imputation with estochastic regression

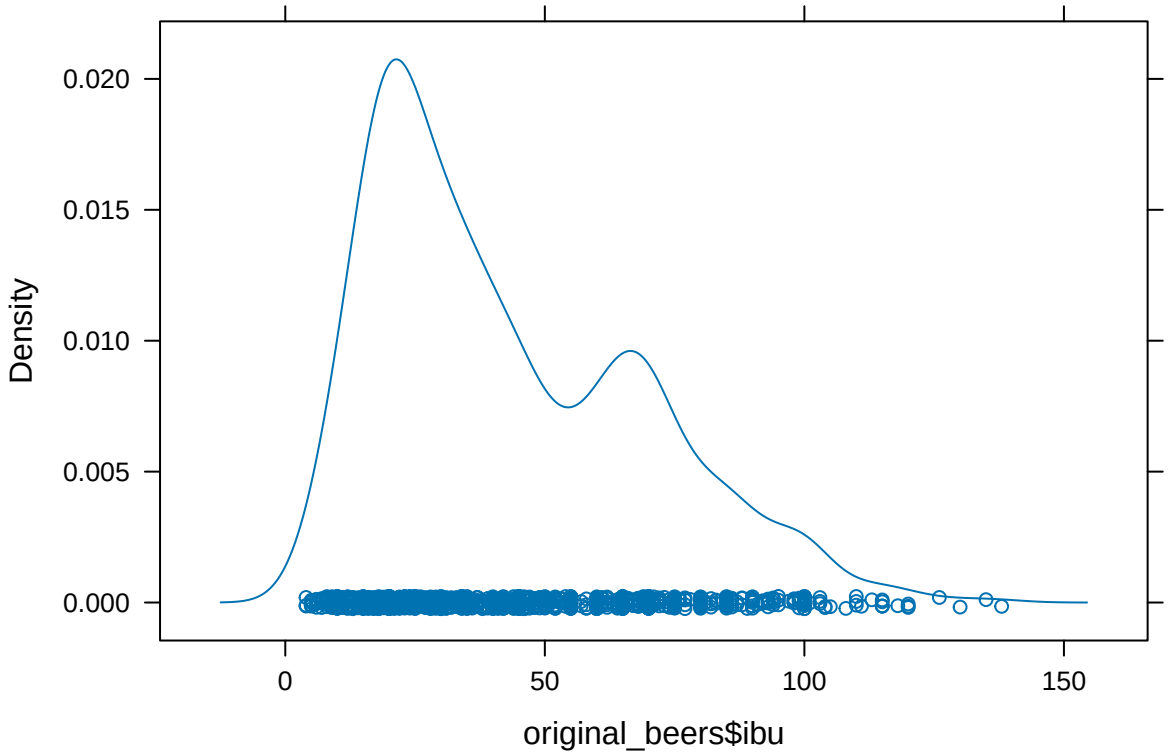
```
par(mfrow = c(2, 2))
densityplot(original_beers$abv)
```



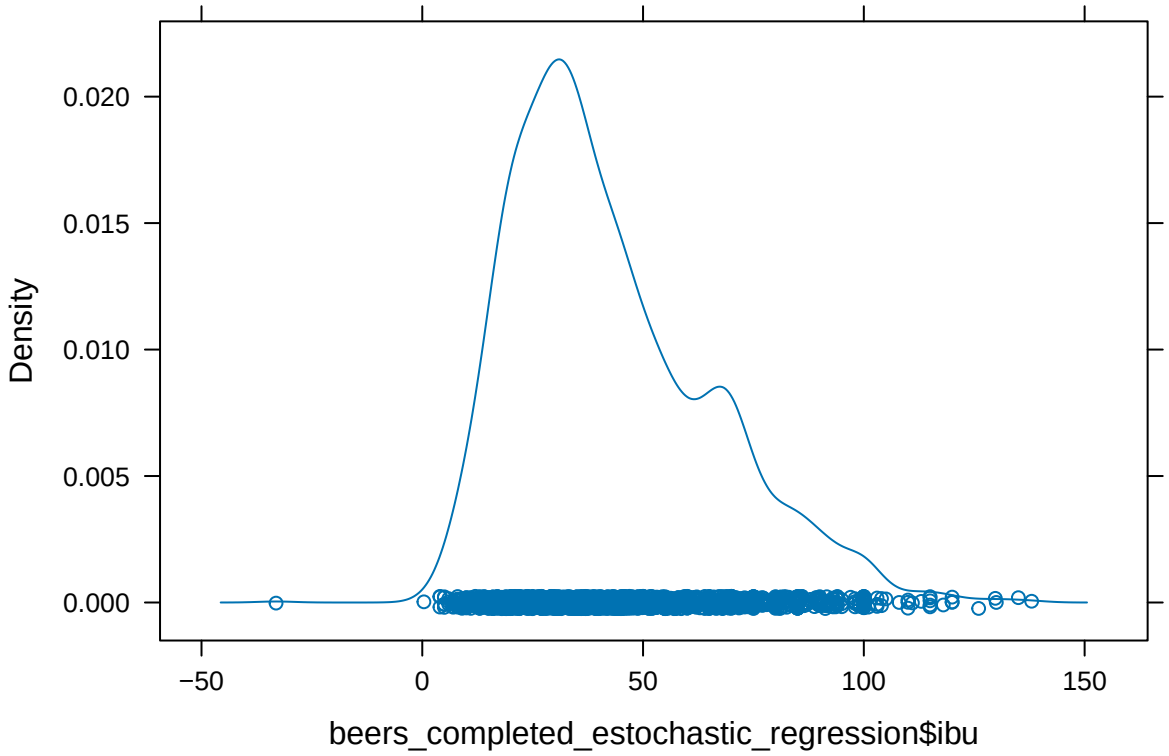

```
densityplot(beers_completed_estochastic_regression$abv)
```




```
densityplot(original_beers$ibu)
```




```
densityplot(beers_completed_estochastic_regression$ibu)
```



And we obtained a rather more coherent completed data, in comparison with the original data.

Some exploratory data analysis

To check the states with the more breweries in the US by the year of the data.

```
original_breweries |> group_by(state) |> summarise(breweries = n()) |> arrange(desc(breweries)) |> slic
```

```
## # A tibble: 10 x 2
##    state breweries
##    <chr>     <int>
##  1 " CO"        47
##  2 " CA"        39
##  3 " MI"        32
##  4 " OR"        29
##  5 " TX"        28
##  6 " PA"        25
##  7 " MA"        23
##  8 " WA"        23
##  9 " IN"        22
## 10 " WI"        20
```

What about the cities with more breweries?

```
original_breweries |> group_by(state, city) |> summarise(breweries = n()) |> arrange(desc(breweries))
```

```
## `summarise()` has grouped output by 'state'. You can override using the
## `.groups` argument.
```

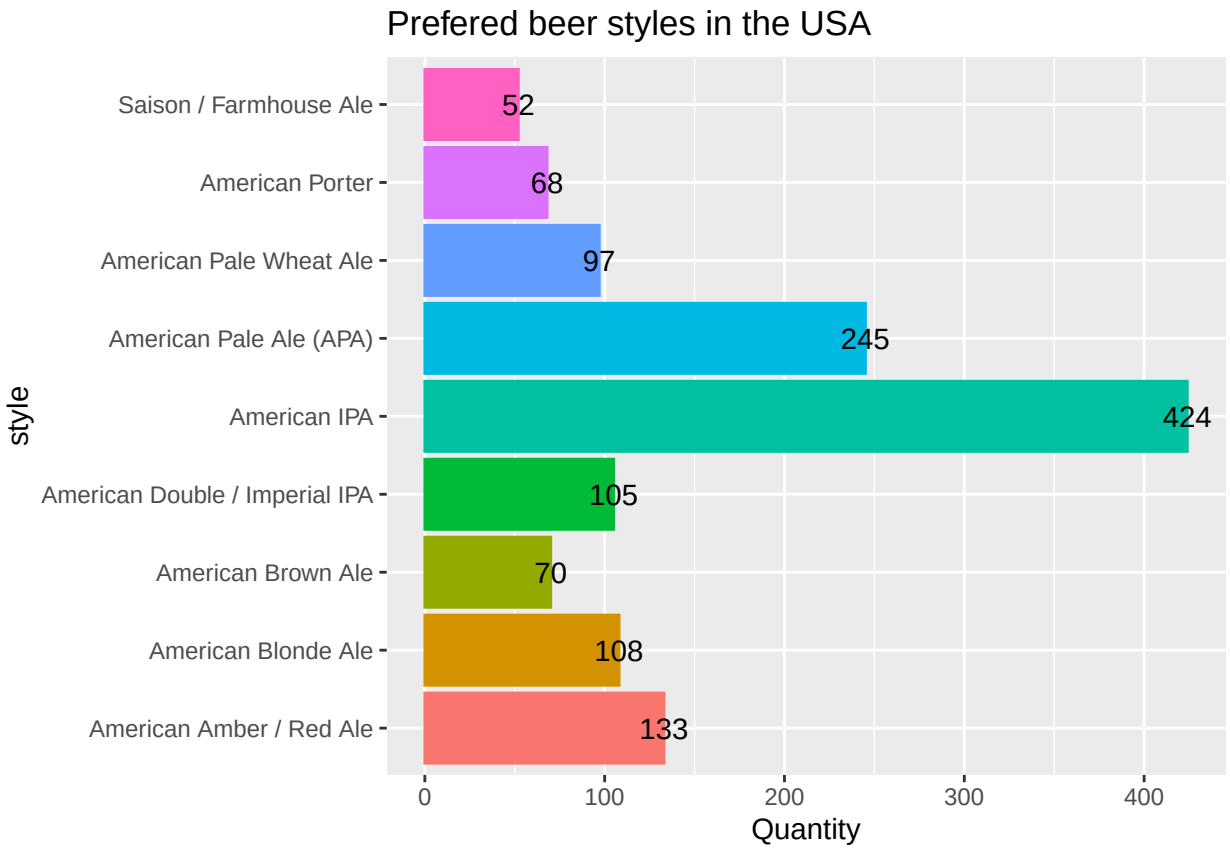
```
## # A tibble: 401 x 3
## # Groups:   state [51]
##    state city          breweries
##    <chr> <chr>             <int>
##  1 " OR" Portland             11
##  2 " CO" Boulder               9
##  3 " IL" Chicago               9
##  4 " WA" Seattle               9
##  5 " CA" San Diego             8
##  6 " CO" Denver                8
##  7 " TX" Austin                8
##  8 " ME" Portland              6
##  9 " OR" Bend                  6
## 10 " CA" San Francisco         5
## # i 391 more rows
```

Let's check the beer's tastes preference in the state.

```
total_styles <- length(beers_completed_estochastic_regression$style)
beers_completed_estochastic_regression |> group_by(style) |> summarise(Quantity = n(), Percentage = Quan
```

```
## # A tibble: 9 x 3
##   style                          Quantity Percentage
##   <chr>                             <int>      <dbl>
## 1 American IPA                        424      17.6
## 2 American Pale Ale (APA)             245      10.2
## 3 American Amber / Red Ale            133       5.52
## 4 American Blonde Ale                 108       4.48
## 5 American Double / Imperial IPA      105       4.36
## 6 American Pale Wheat Ale              97       4.02
## 7 American Brown Ale                   70       2.90
## 8 American Porter                      68       2.82
## 9 Saison / Farmhouse Ale               52       2.16
```

```
beers_completed_estochastic_regression |> group_by(style) |> summarise(Quantity = n()) |> arrange(desc(
  geom_col(orientation = 'y') +
  theme(legend.position = 'none') +
  ggtitle('Prefered beer styles in the USA') +
  geom_text(aes(label = Quantity), colour = 'black')
```


Prefered beer styles in the USA
Saison / Farmhouse Ale
American Porter
American Pale Wheat Ale
American Pale Ale (APA)
American IPA
American Double / Imperial IPA
American Brown Ale
American Blonde Ale
American Amber / Red Ale
52
68
97
245
424
105
70
108
133
style
0
100
200
300
400
Quantity