Imputación de datos faltantes en base de datos de cervecerías

Daniel Navarro

2025-08-03

Objetivo

Para revisar una imputación de variables múltiples con equaciones en cadena tomamos una base de datos de Kaggle https://www.kaggle.com/datasets/nickhould/craft-cans/data que contiene dos bases de datos con 2410 diferentes cervezas y 558 en la base de datos de fabricantes en los Estados Unidos.

Las bases de datos están incompletas, faltan 1067 datos distribuidos en diferentes campos de las cervezas y 0 en los de fabricantes.

No se puede asumir descartar 1067 lineas de un total de 2410 y se debe imputar estos datos, pero, ¿Cómo están distribuidos los datos faltantes?

Se describe a continuación las bases de datos, abv es el contenido de alcohol por volumen, ibu es la unidad internacional de sabor amargo, el resto de las variables están bien identificadas:

```
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
##
```

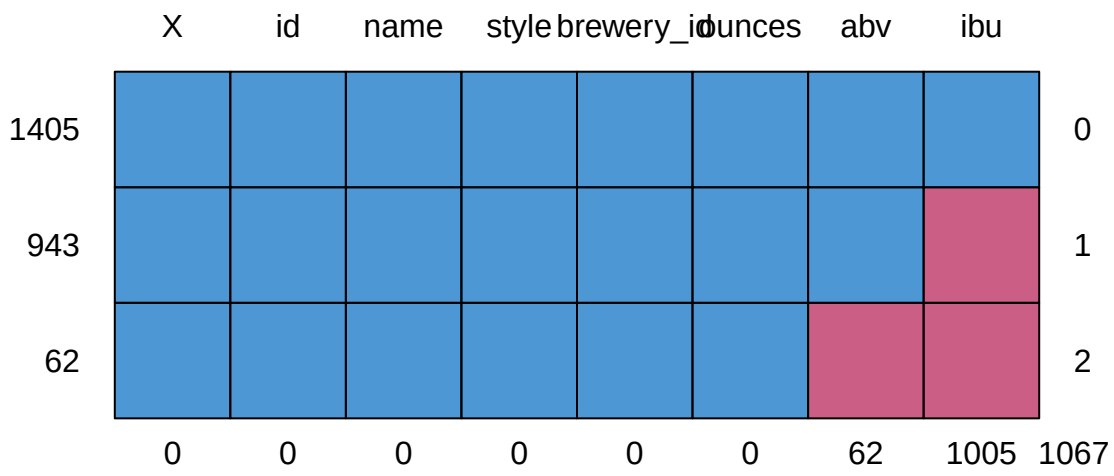
```
summary(original_breweries)
```

```
##        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
```

Imputación de datos

Revisamos cómo están distribuidos los datos faltantes en la base de datos de cervezas, beers.

```
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
```

Solo faltan datos en las variables abv e ibu, 943 líneas carecen solo de ibu y 62 de ambas variables, abv e ibu.

Imputación simple utilizando la media

Esta es la práctica común de usar la media de los datos existentes en los datos faltantes.

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

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

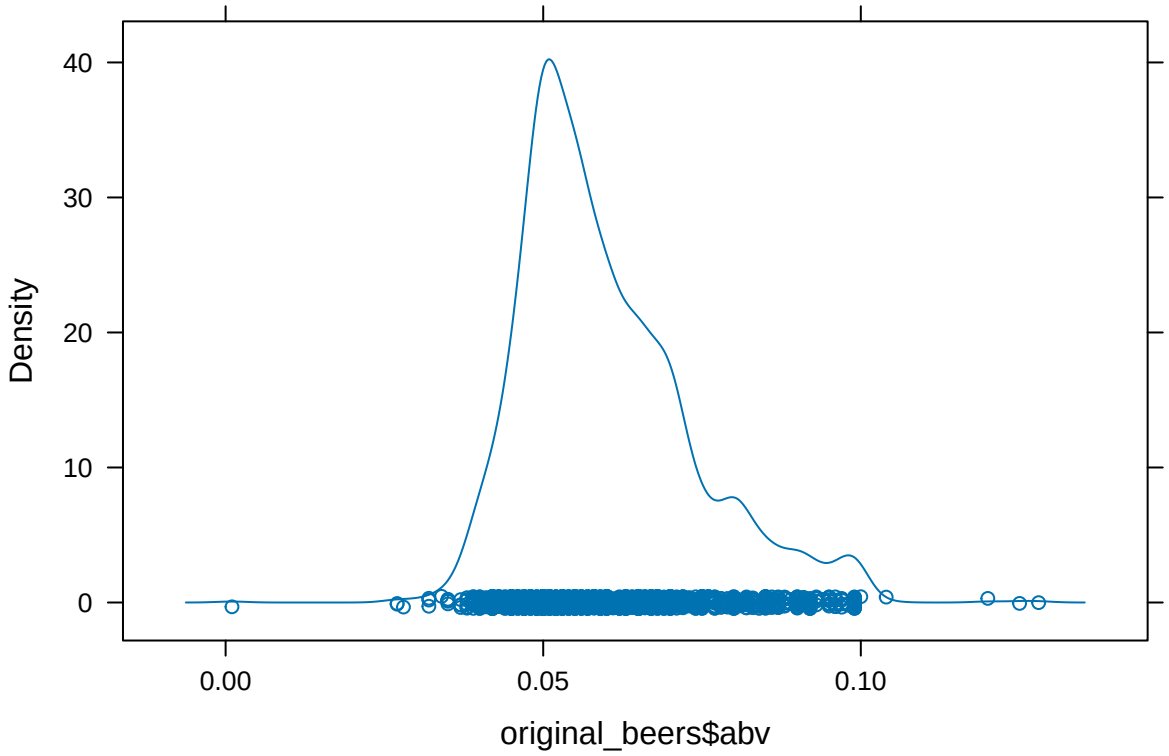
```
beers_completed_mean <- complete(imp_mean)
summary(beers_completed_mean)
```

```
##        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
```

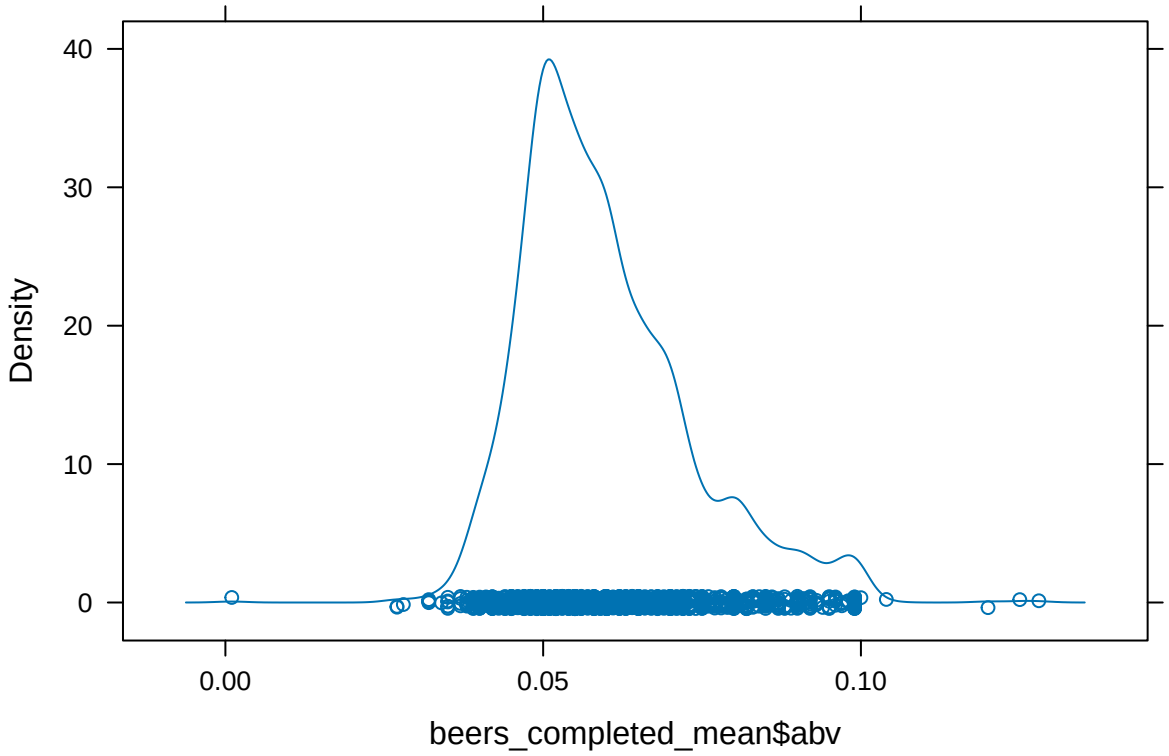
Al completar los datos faltantes con la media, se ha completado en una acción todos los datos faltantes en la base de datos.

Al comparar las densidades de ambas variables, antes y después de imputar los datos con la media, abv es consistente con los datos originales, sin embargo ibu está demasiado centrada. Este es un efecto del uso del promedio en casi la mitad de todas las líneas de la variable ibu, así que la media no es una forma balanceada de completar estos datos.

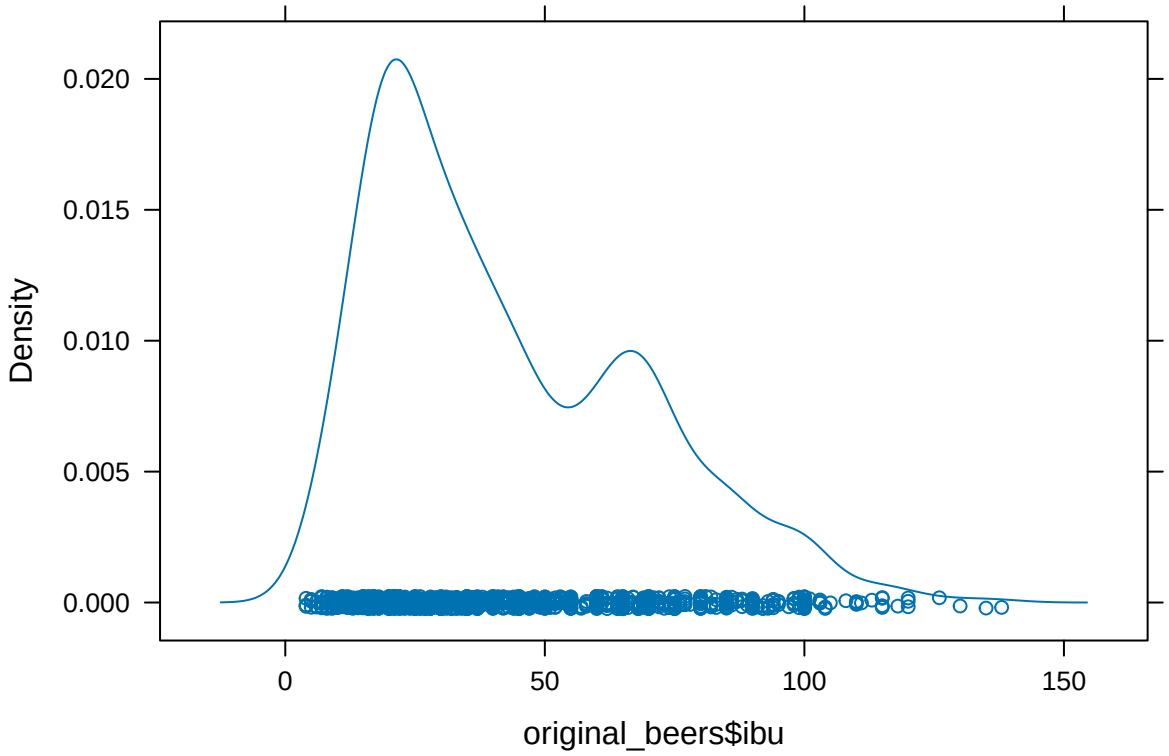
```
densityplot(original_beers$abv)
```



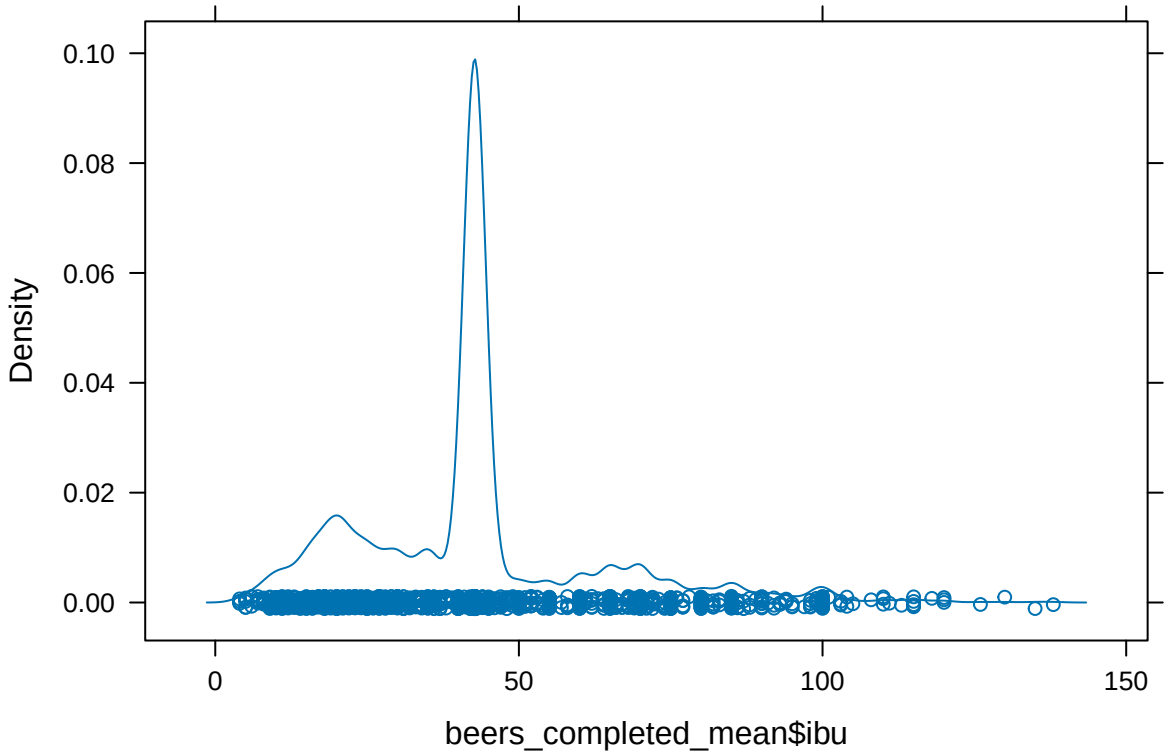

```
densityplot(beers_completed_mean$abv)
```




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



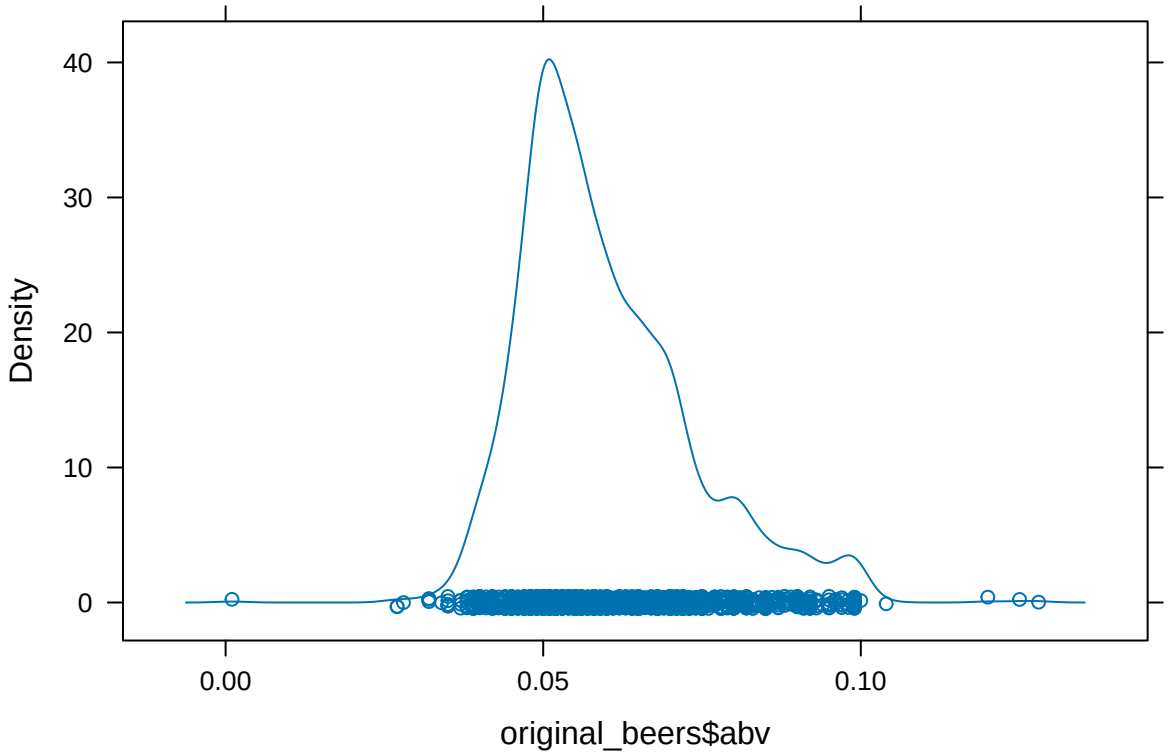

```
densityplot(beers_completed_mean$ibu)
```



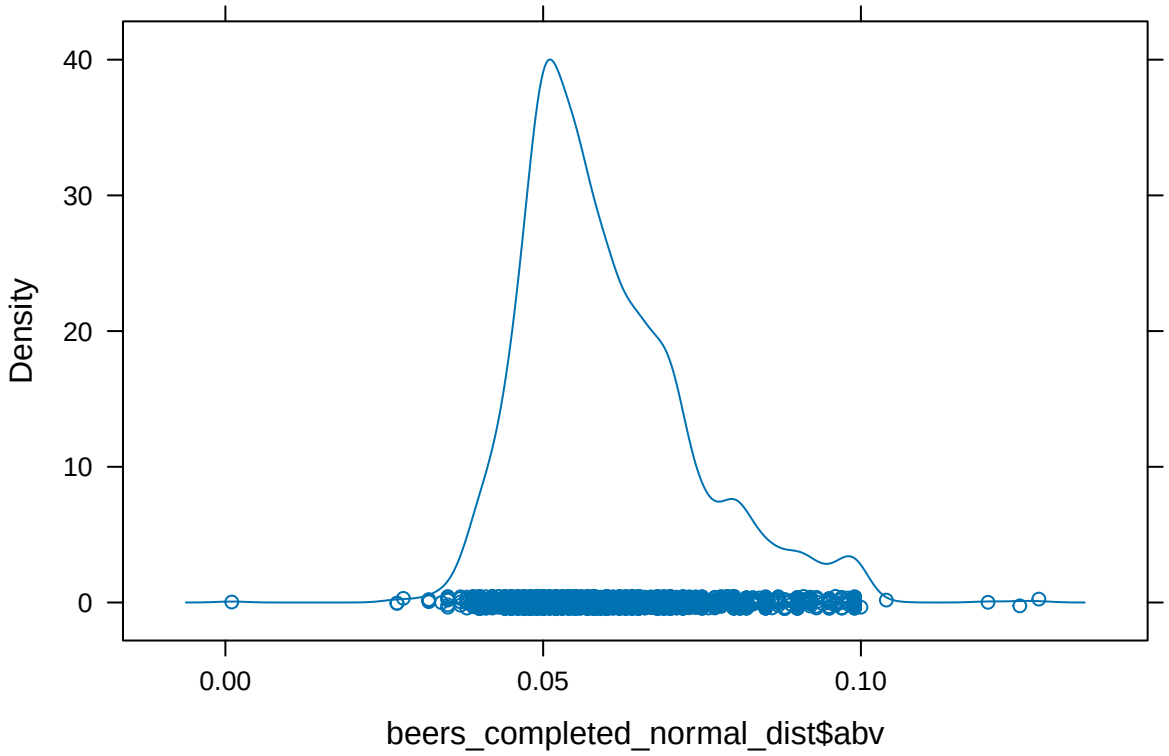
Probaremos otro método de imputación:

Imputación de datos con distribución normal

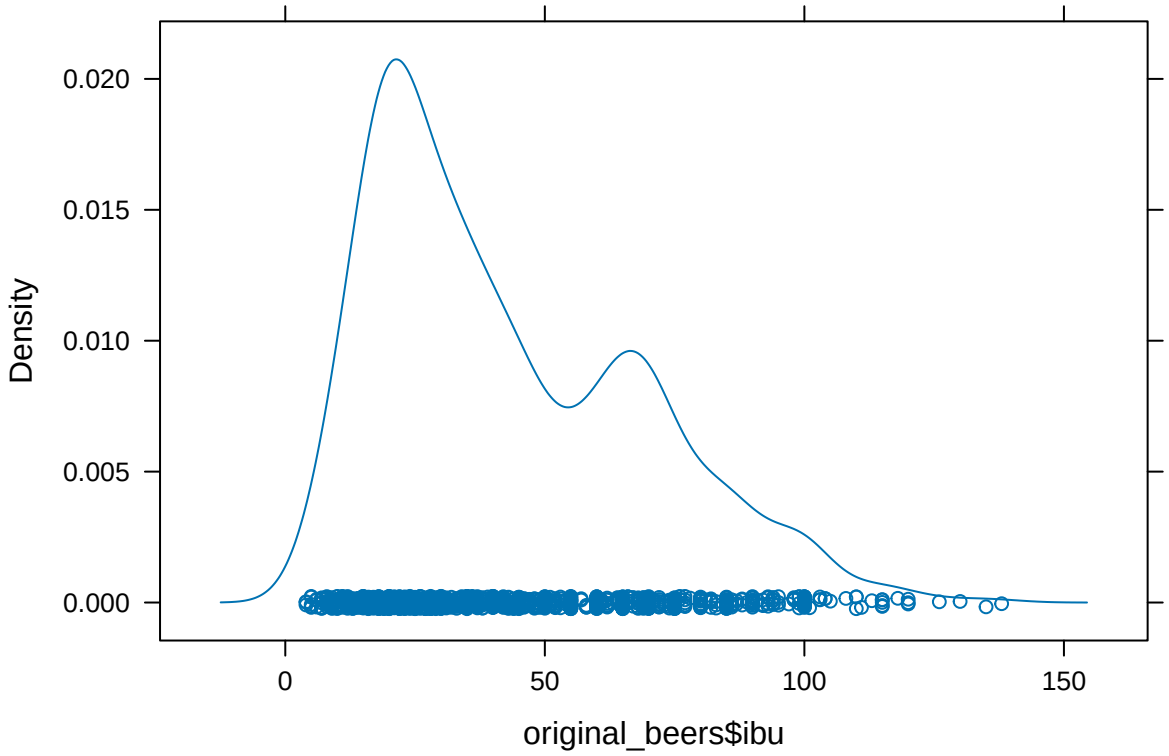
```
densityplot(original_beers$abv)
```



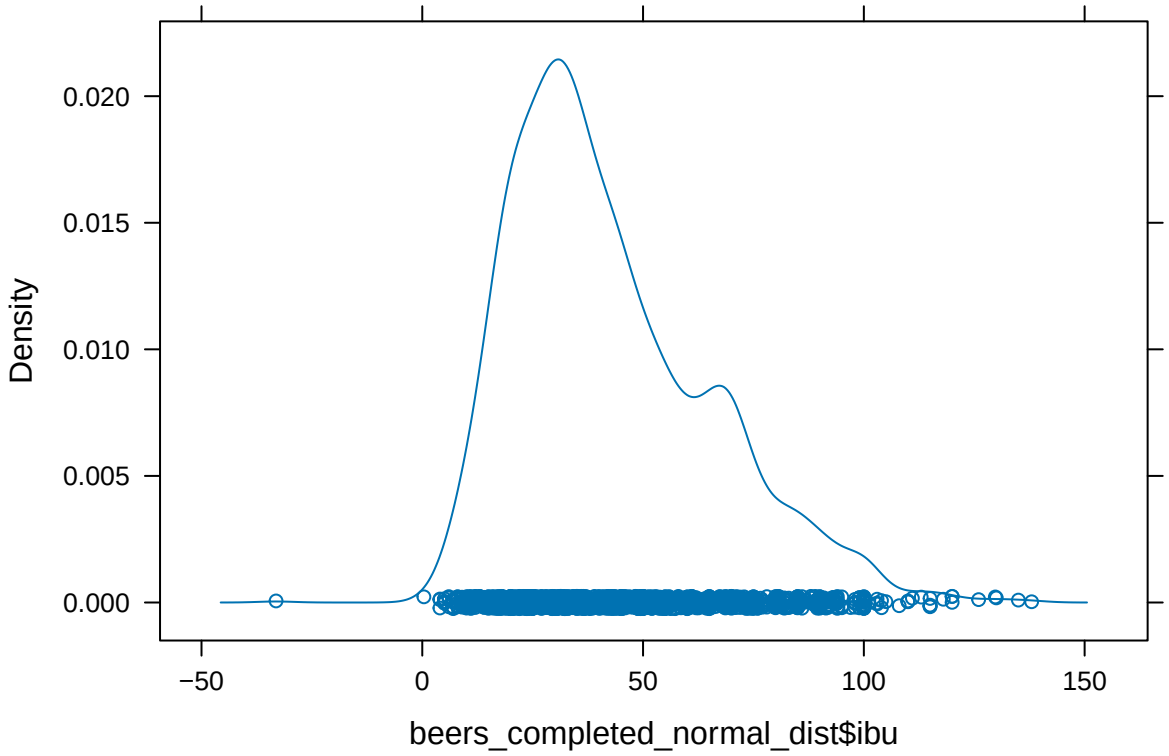

```
densityplot(beers_completed_normal_dist$abv)
```




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




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



Ahora obtenemos datos completados con una distribución coherente con los datos originales.

Sin embargo, para satisfacer la curiosidad y por práctica podemos probar otro método, esta vez:

Imputación de datos con regresión estocástica

```
beers_completed_estochastic_regression <- original_beers
imp_normal <- mice(beers_completed_estochastic_regression, method = 'norm.predict', m = 1, maxit = 1)
```

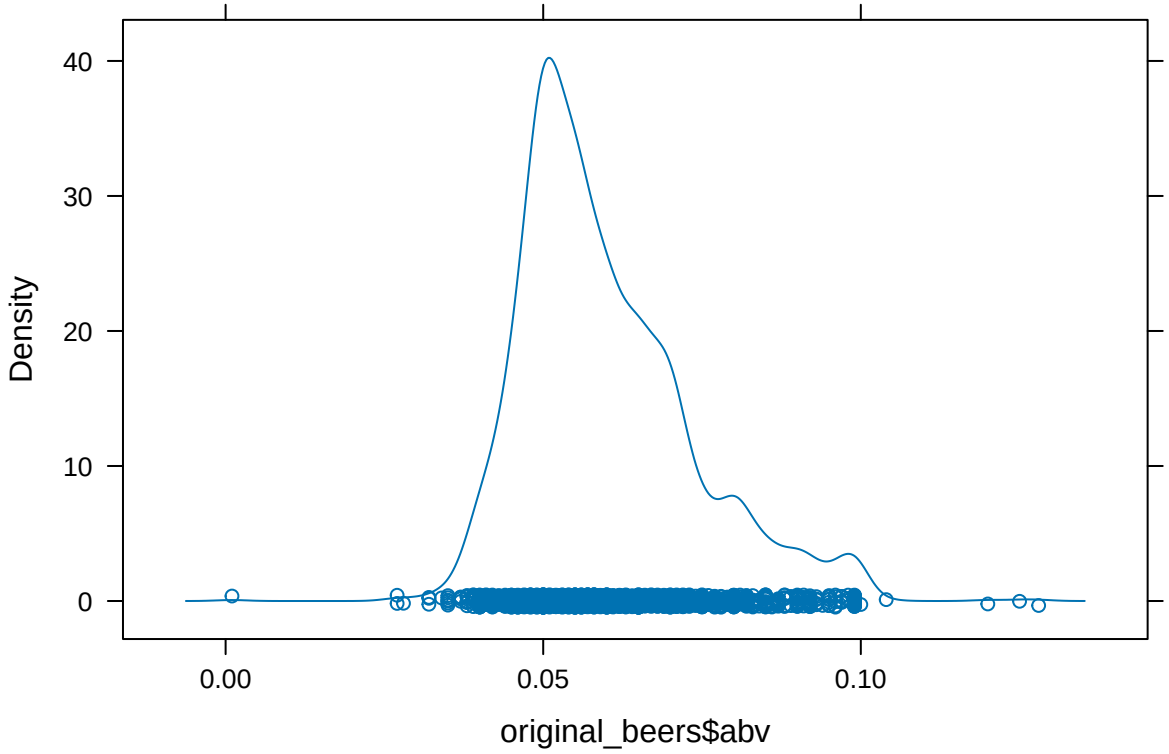
```
##
##  iter imp variable
##   1   1  abv  ibu
```

```
beers_completed_estochastic_regression <- complete(imp_normal)
summary(beers_completed_estochastic_regression)
```

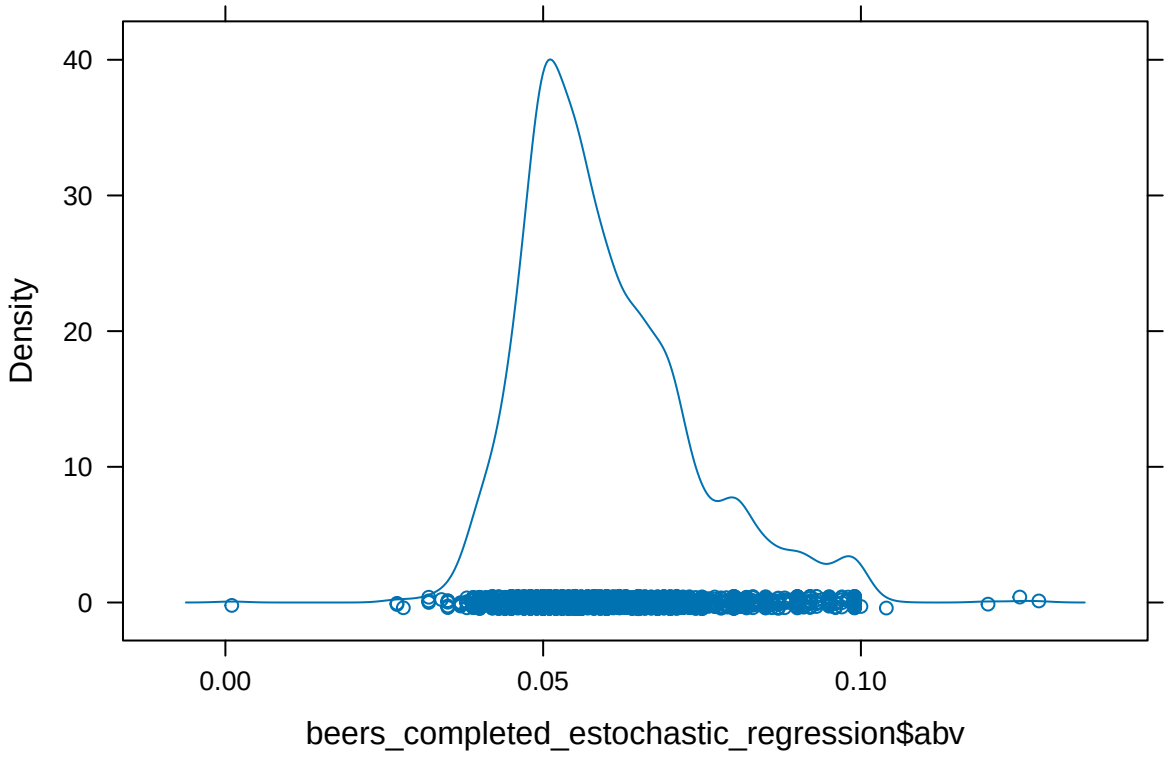
```
##        X              abv               ibu              id
##  Min.   :   0.0   Min.   :0.00100   Min.   :-33.11   Min.   :   1.0
##  1st Qu.: 602.2   1st Qu.:0.05000   1st Qu.: 25.63   1st Qu.: 808.2
##  Median :1204.5   Median :0.05643   Median : 37.00   Median :1453.5
##  Mean   :1204.5   Mean   :0.05977   Mean   : 42.54   Mean   :1431.1
##  3rd Qu.:1806.8   3rd Qu.:0.06700   3rd Qu.: 55.07   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
```

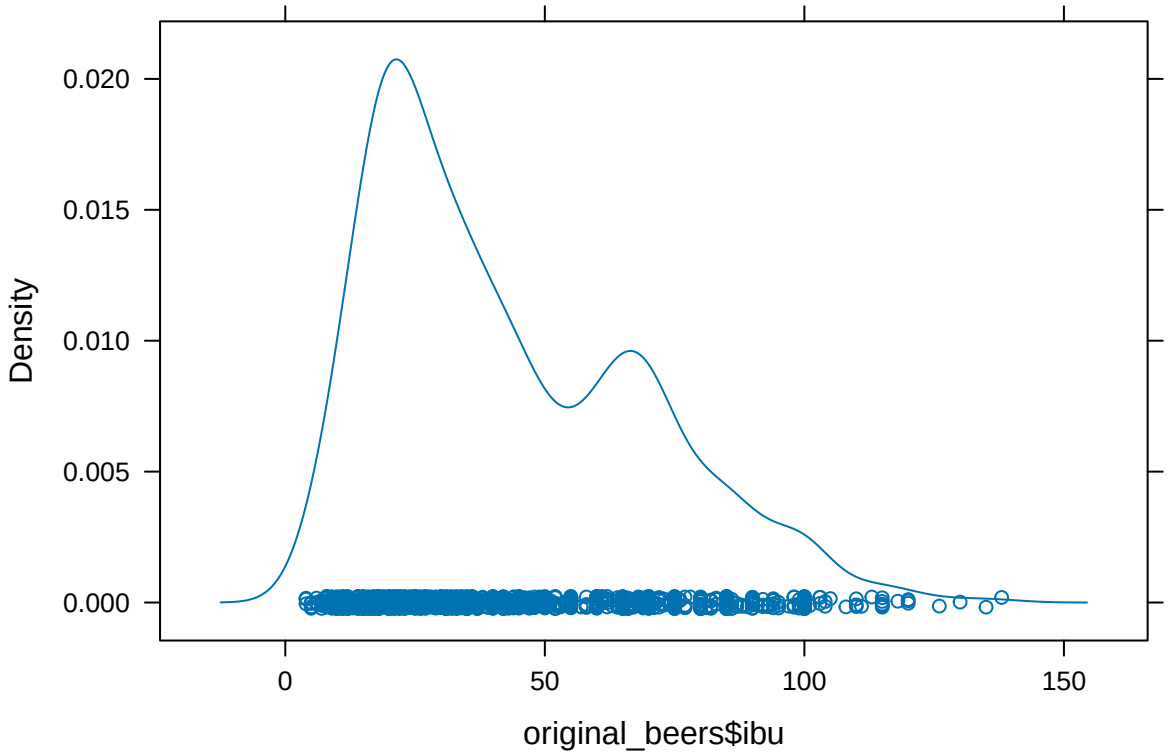
```
densityplot(original_beers$abv)
```



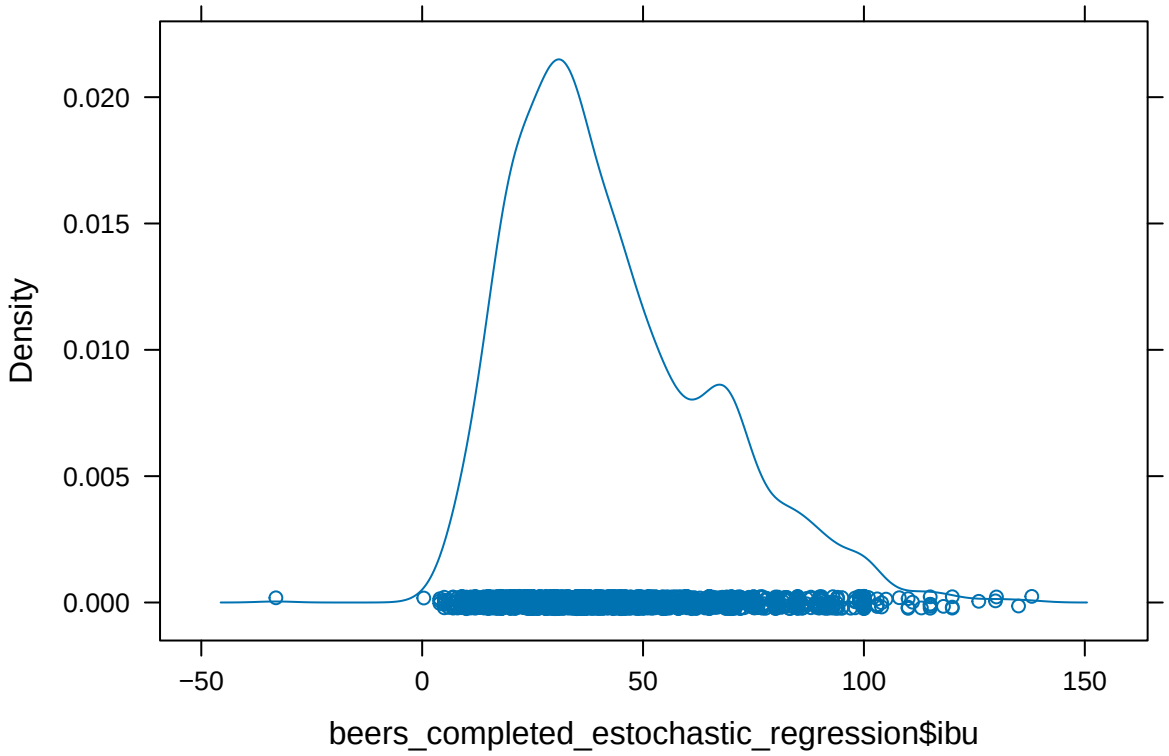

```
densityplot(beers_completed_estochastic_regression$abv)
```




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




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



Y obtenemos unos datos mucho más coherentes al compararlos con los datos originales.

Análisis exploratorio de los datos

Estados con más productores de cerveza en los Estados Unidos a la fecha de la base de datos.

```
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
```

¿Qué hay de las ciudades con más productores de cerveza?

```
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.
```

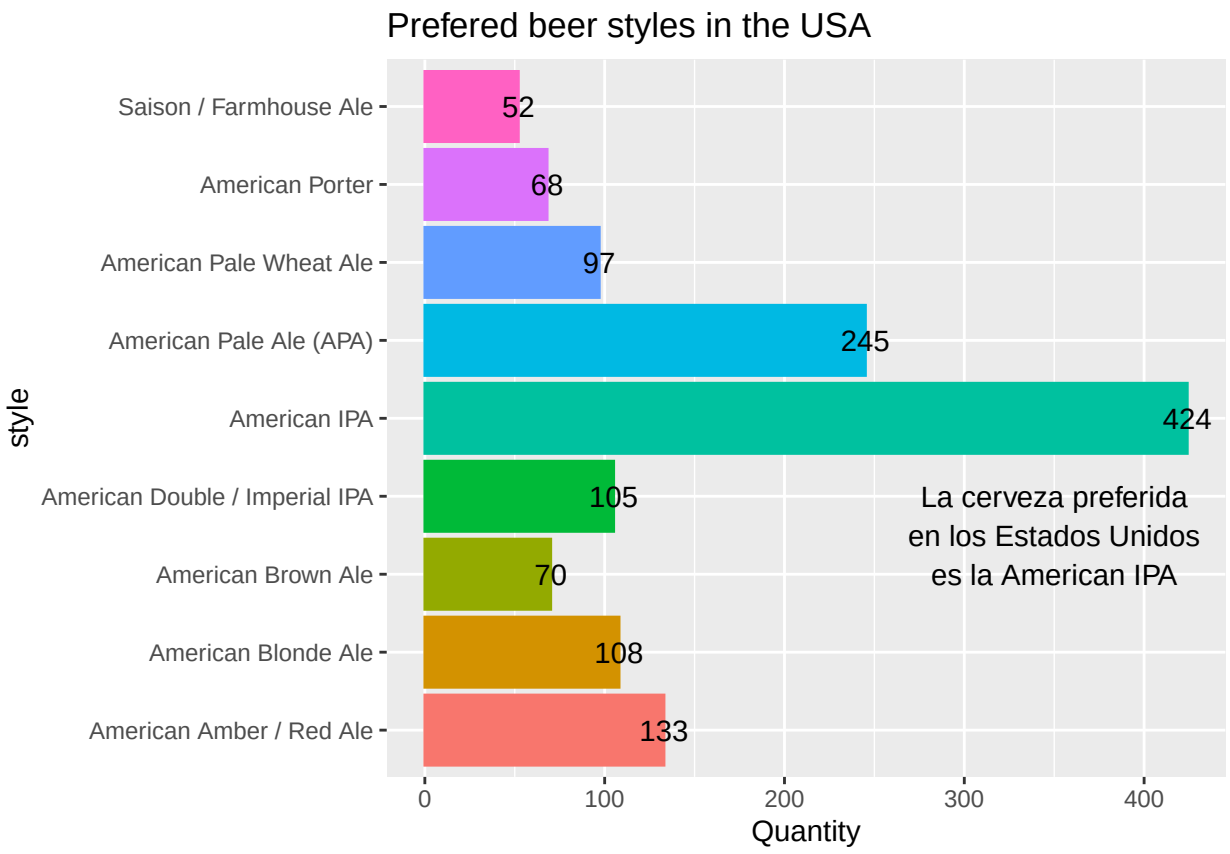
```
## # A tibble: 401 x 3
## # Groups:   state [51]
##    state city          breweries
##    <chr> <chr>             <int>
##  1 " OR" Portland             11
##  2 " CO" Boulder               9
##  3 " IL" Chicago               9
##  4 " WA" Seattle               9
##  5 " CA" San Diego             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
```

Veamos las cervezas preferidas en los Estados Unidos.

```
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') +
  annotate('text', x = 350, y = 4, label = 'La cerveza preferida') +
  annotate('text', x = 350, y = 3.5, label = 'en los Estados Unidos') +
  annotate('text', x = 350, y = 3, label = 'es la American IPA')
```


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
La cerveza preferida
en los Estados Unidos
es la American IPA