DATA 606 Data Project Proposal

Leticia Salazar

Contents

Data Preparation

```
# load packages / libraries
library("tidyverse")
```

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

## v ggplot2 3.3.5     v purrr   0.3.4
## v tibble  3.1.4     v dplyr   1.0.7
## v tidyr   1.1.3     v stringr 1.4.0
## v readr   2.0.1     v forcats 0.5.1

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

```
library("tidyr")
library("dplyr")
library("ggplot2")
```

```
# load data
drinks <- read.csv("https://raw.githubusercontent.com/letisalba/Data-606/main/Project/drinks.csv", heade
```

```
glimpse(drinks)
```

```
## Rows: 193
## Columns: 5
## $ country                      <chr> "Afghanistan", "Albania", "Algeria", "And~
## $ beer_servings                <int> 0, 89, 25, 245, 217, 102, 193, 21, 261, 2~
## $ spirit_servings              <int> 0, 132, 0, 138, 57, 128, 25, 179, 72, 75,~
## $ wine_servings                <int> 0, 54, 14, 312, 45, 45, 221, 11, 212, 191~
## $ total_litres_of_pure_alcohol <dbl> 0.0, 4.9, 0.7, 12.4, 5.9, 4.9, 8.3, 3.8, ~
```

```r
# Get column names
names(drinks)
```

```
## [1] "country"                      "beer_servings"
## [3] "spirit_servings"              "wine_servings"
## [5] "total_litres_of_pure_alcohol"
```

```r
# Rename columns
colnames(drinks) <- c("Country", "Beer_Servings", "Spirit_Servings", "Wine_Servings", "Total_Litres_Pur
```

Research question

You should phrase your research question in a way that matches up with the scope of inference your data set allows for.

In general, is there a significant difference in the preferred type of alcohol?

Cases

What are the cases, and how many are there?

Each case represents a country around the world along with their beer, spirits and/or wine number of servings, as well as the total liters of pure alcohol. There are 193 total observations in this data set.

Data collection

Describe the method of data collection.

The data was collected from FiveThirtyEight's article called "Dear Mona Followup: Where Do People Drink The Most Beer, Wine And Spirits?" This data was collected by World Health Organisation, Global Information System on Alcohol and Health (GISAH), 2010.

Type of study

What type of study is this (observational/experiment)?

This is an observational study.

Data Source

If you collected the data, state self-collected. If not, provide a citation/link.

World Health Organization. (n.d.). Global information system on alcohol and health. World Health Organization. Retrieved October 19, 2021, from https://www.who.int/data/gho/data/themes/global-information-system-on-alcohol-and-health.

Dependent Variable

What is the response variable? Is it quantitative or qualitative?

The dependent variable is alcohol consumption and it it quantitative.

Independent Variable

What is the independent variable? Is it quantitative or qualitative?

The independent variables are country and types of alcohol and they are qualitative.

Relevant summary statistics

Provide summary statistics for each the variables. Also include appropriate visualizations related to your research question (e.g. scatter plot, boxplots, etc). This step requires the use of R, hence a code chunk is provided below. Insert more code chunks as needed.

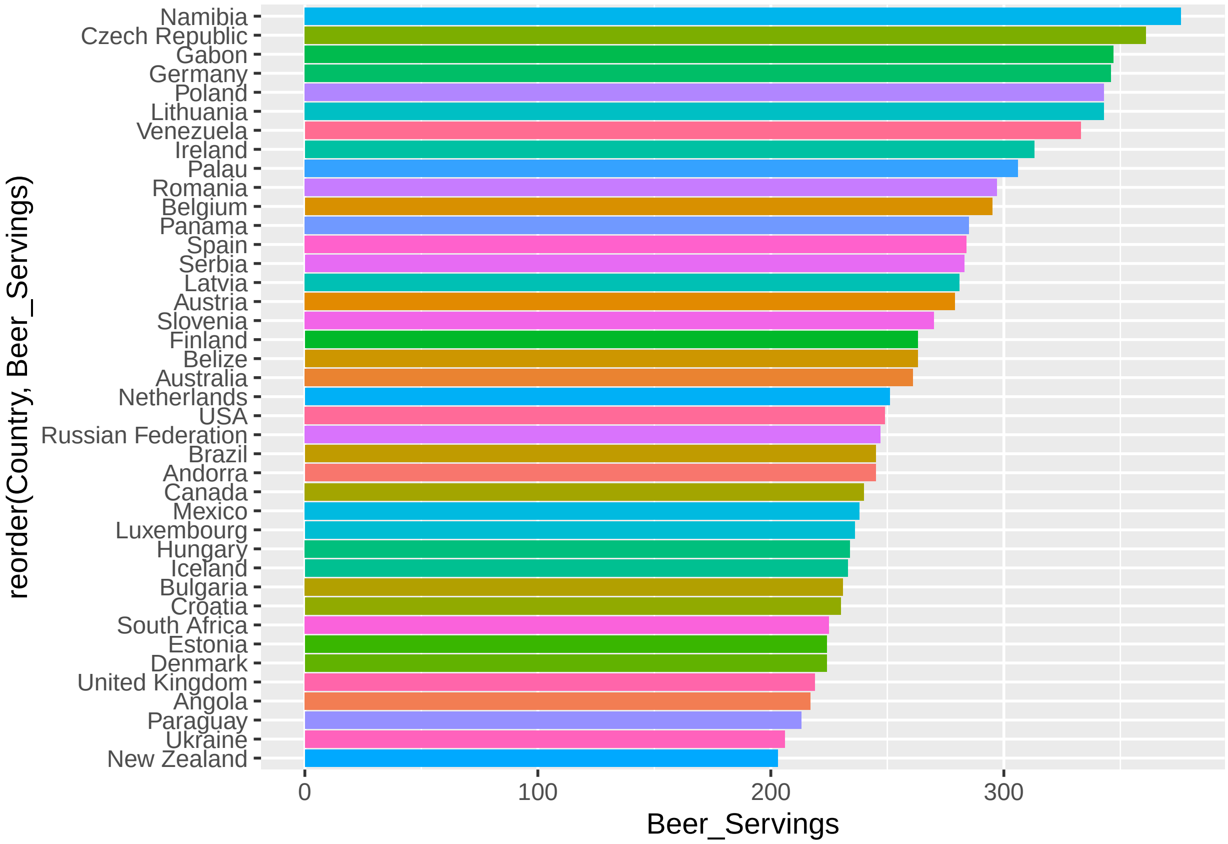
```
summary(drinks$Country)
```

```
##    Length     Class      Mode
##       193 character character
```

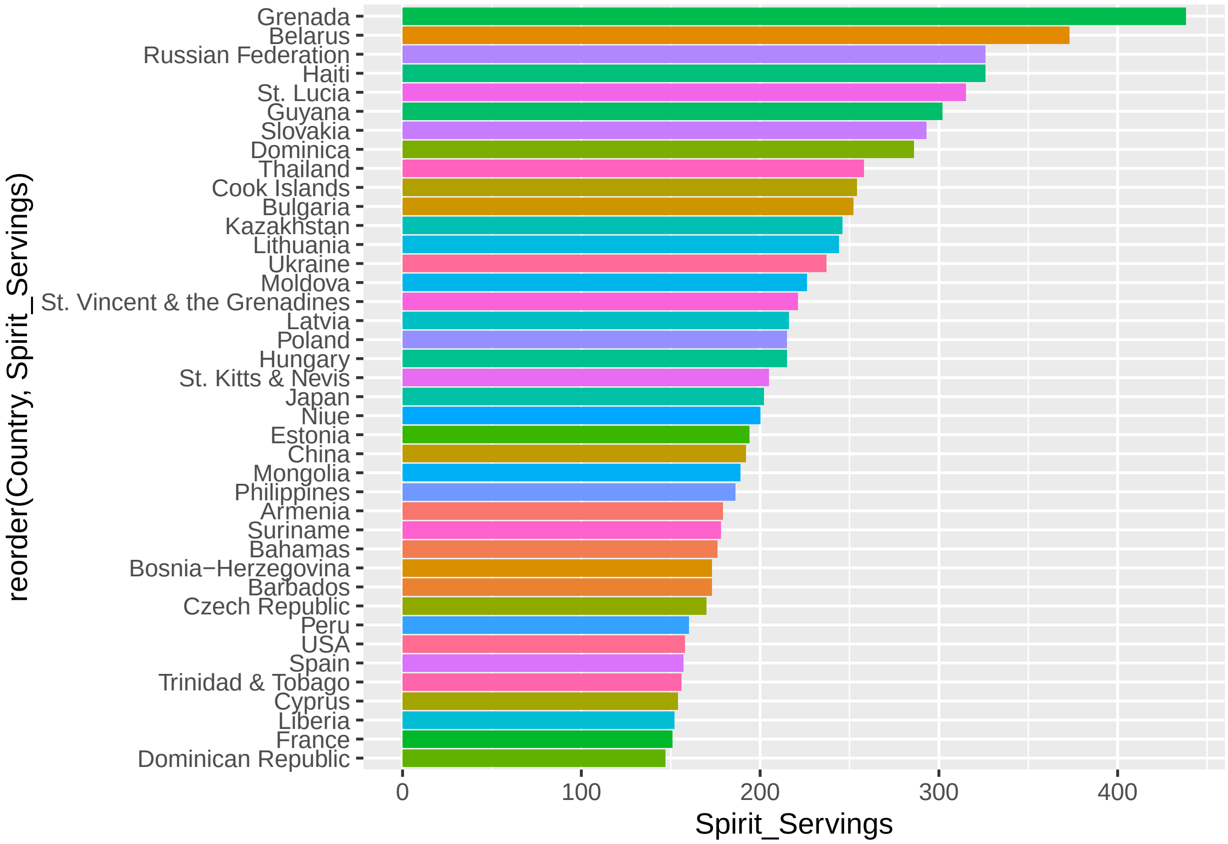
```
summary(drinks$Beer_Servings)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     0.0    20.0    76.0   106.2   188.0   376.0
```

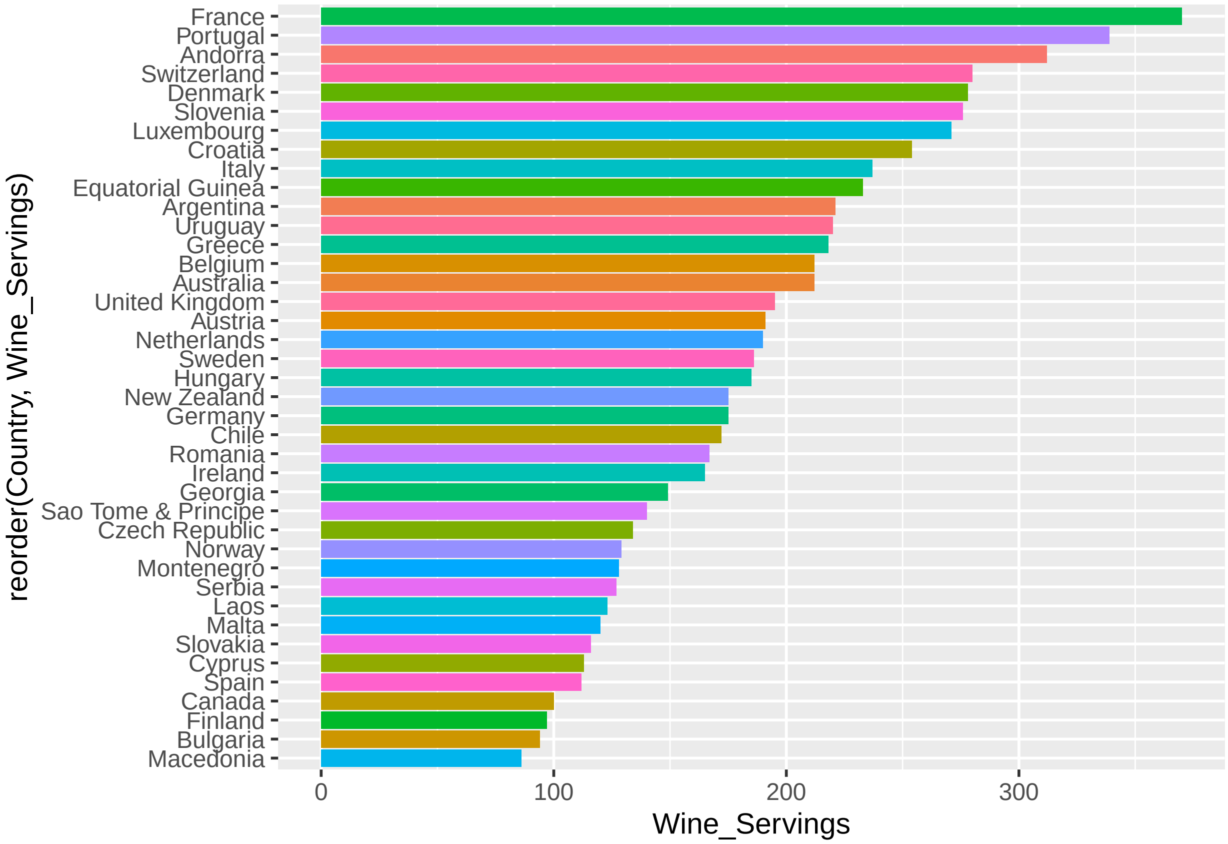
```
# Beer servings by Country
drinks %>%
  arrange(desc(Beer_Servings)) %>%
  head(40) %>%
  ggplot(aes(y = reorder(Country, Beer_Servings),
             x = Beer_Servings,
                  fill= Country))+
  geom_col()+
  theme(legend.position = "none")
```



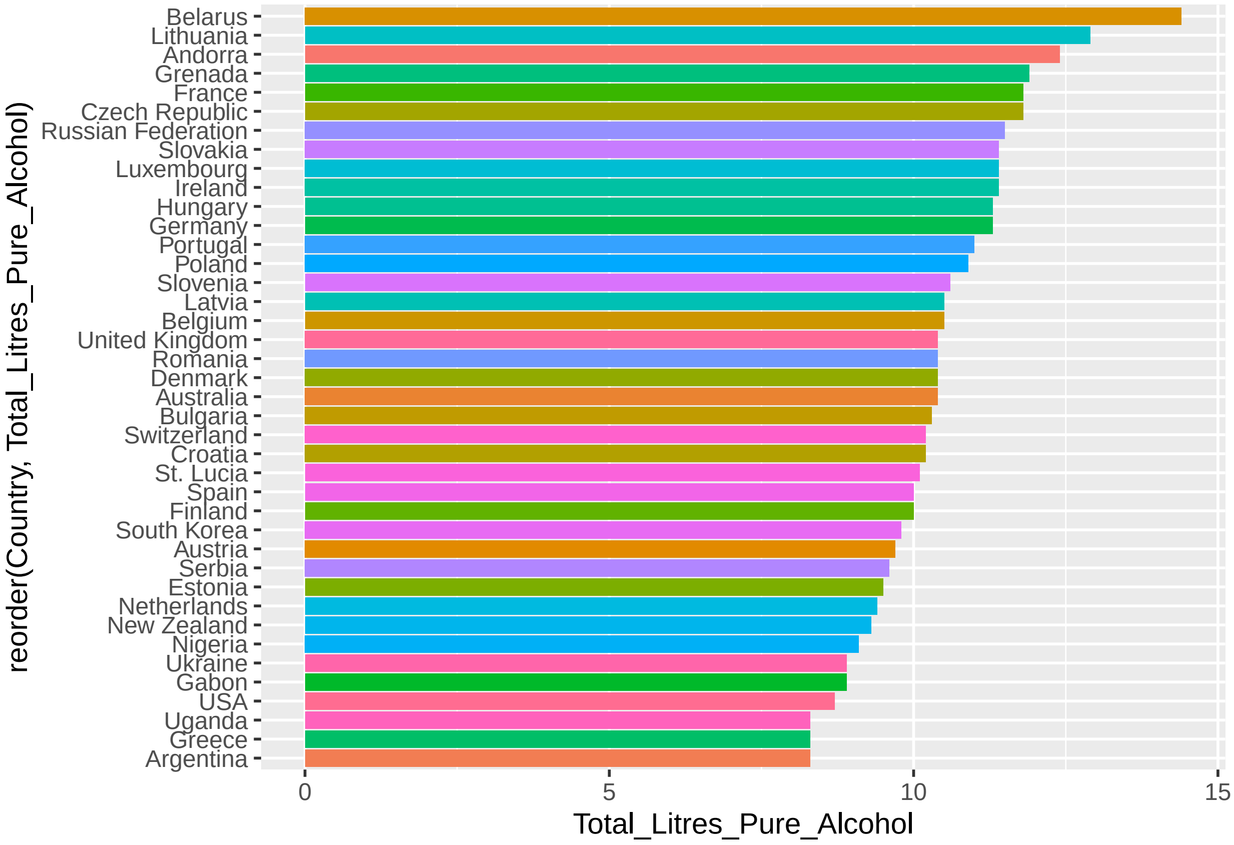
```
# Spirits serving by Country top 40
drinks %>%
  arrange(desc(Spirit_Servings)) %>%
  head(40) %>%
  ggplot(aes(y = reorder(Country, Spirit_Servings),
             x = Spirit_Servings,
                  fill= Country))+
  geom_col()+
  theme(legend.position = "none")
```



```
# Wine servings by Country top 40
drinks %>%
  arrange(desc(Wine_Servings)) %>%
  head(40) %>%
  ggplot(aes(y = reorder(Country, Wine_Servings),
             x = Wine_Servings,
                  fill= Country))+
  geom_col()+
  theme(legend.position = "none")
```

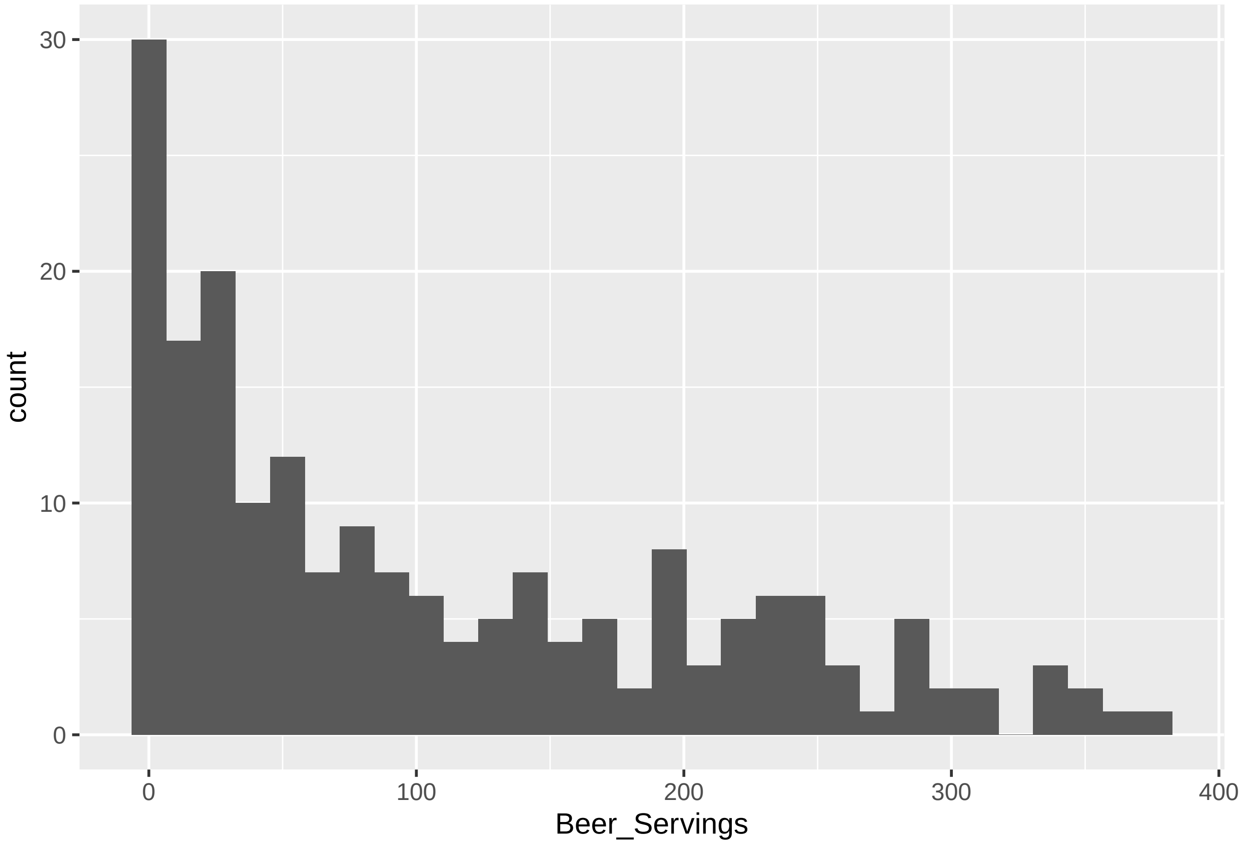


```
# Total alcohol in litres by Country top 40
drinks %>%
  arrange(desc(Total_Litres_Pure_Alcohol)) %>%
  head(40) %>%
  ggplot(aes(y = reorder(Country, Total_Litres_Pure_Alcohol),
             x = Total_Litres_Pure_Alcohol,
                fill= Country))+
  geom_col()+
  theme(legend.position = "none")
```



```
drinks %>%
ggplot(aes(x = Beer_Servings)) +
  geom_histogram()
```

```
## ‘stat_bin()‘ using ‘bins = 30‘. Pick better value with ‘binwidth‘.
```

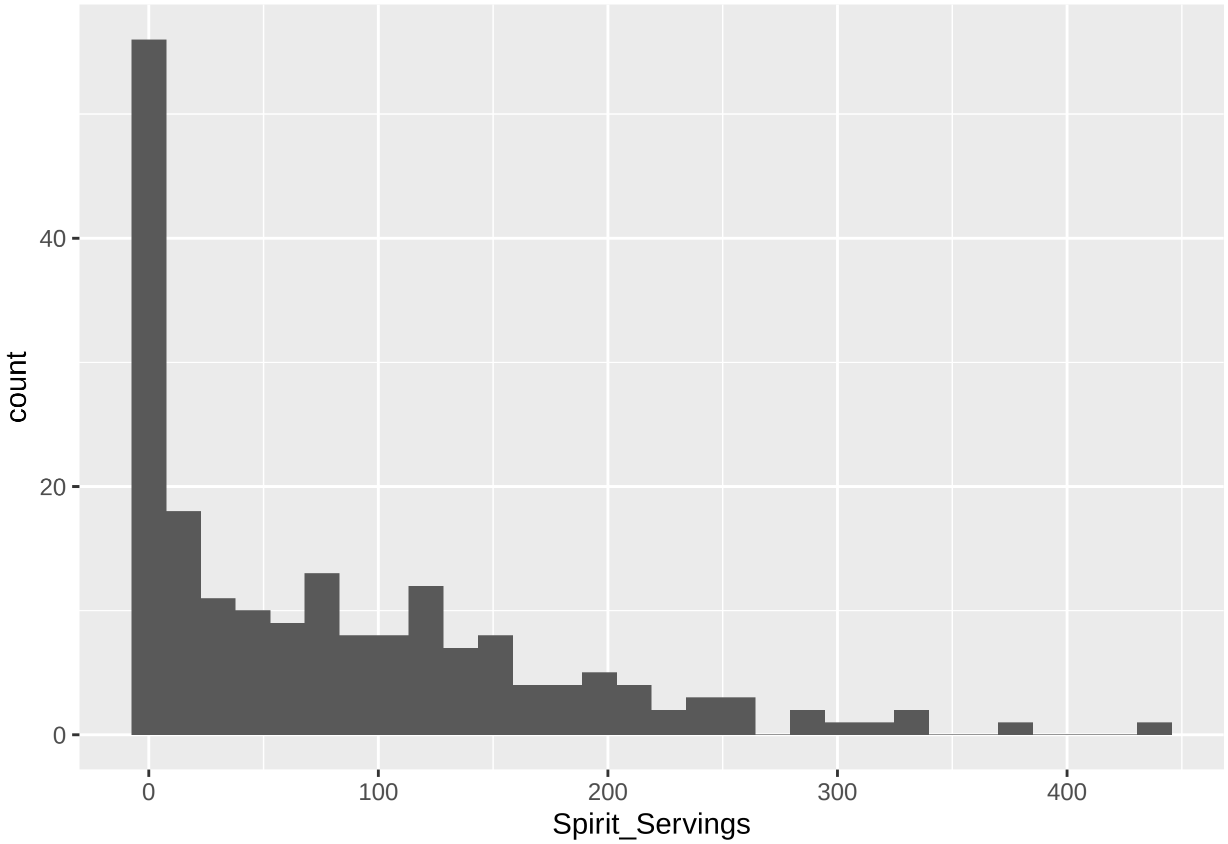


```
summary(drinks$Spirit_Servings)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##    0.00    4.00   56.00   80.99  128.00  438.00 
```

```
drinks %>%
ggplot(aes(x = Spirit_Servings)) +
  geom_histogram()
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

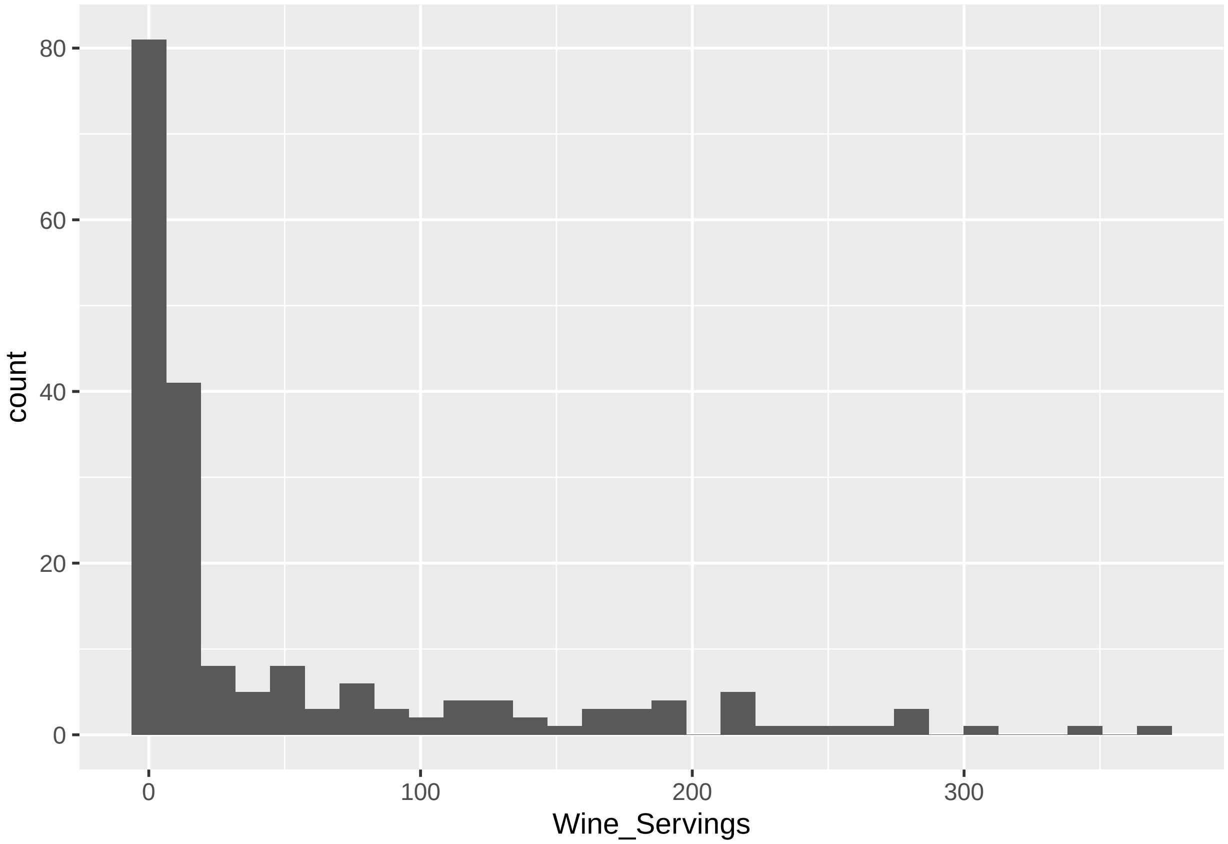


```
summary(drinks$Wine_Servings)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##    0.00    1.00    8.00   49.45   59.00  370.00
```

```
drinks %>%
ggplot(aes(x = Wine_Servings)) +
  geom_histogram()
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

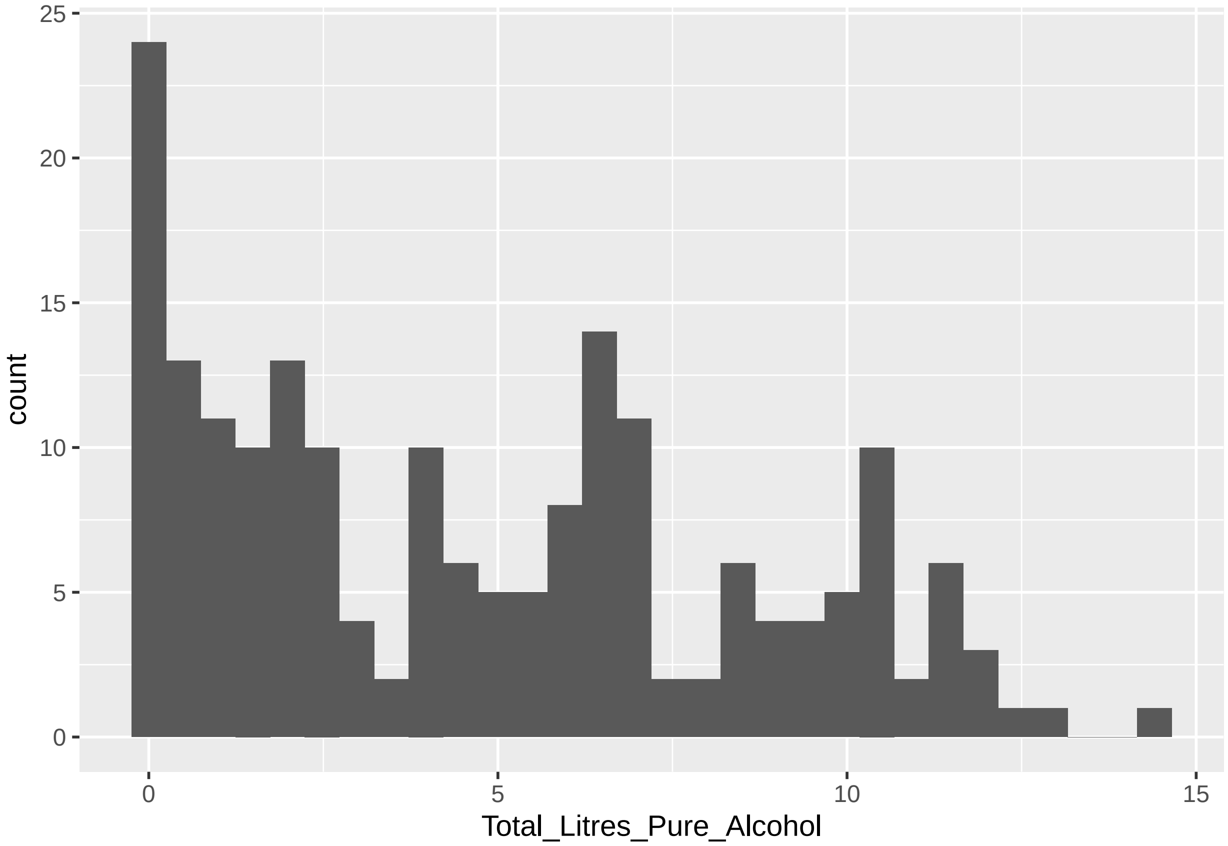


```
summary(drinks$Total_Litres_Pure_Alcohol)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.000   1.300   4.200   4.717   7.200  14.400
```

```
drinks %>%
ggplot(aes(x = Total_Litres_Pure_Alcohol)) +
  geom_histogram()
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```


25
20
15
10
5
0
count
0
5
10
15
Total_Litres_Pure_Alcohol