HW2

2022-09-28

Rules:

Every plot must have a description and explanation on it. If the plot does not have explanation, you will lose points

You have to submit a knitted(PDF document) also. If a PDF document is missing, you will lose points

The dataset describes Airbnb metrics (hotels/house hosting, geographical availability, pricing) in NYC for 2019.

Load the NYC dataset. Check the structure and the summary of the data.

```
nyc <- read.csv("AB_NYC_2019.csv")
str(nyc)
```

```
## 'data.frame':    48895 obs. of  16 variables:
##  $ id                            : int  2539 2595 3647 3831 5022 5099 5121 5178 5203 5238 ...
##  $ name                          : chr  "Clean & quiet apt home by the park" "Skylit Midtown Castle"
##  $ host_id                       : int  2787 2845 4632 4869 7192 7322 7356 8967 7490 7549 ...
##  $ host_name                     : chr  "John" "Jennifer" "Elisabeth" "LisaRoxanne" ...
##  $ neighbourhood_group           : chr  "Brooklyn" "Manhattan" "Manhattan" "Brooklyn" ...
##  $ neighbourhood                 : chr  "Kensington" "Midtown" "Harlem" "Clinton Hill" ...
##  $ latitude                      : num  40.6 40.8 40.8 40.7 40.8 ...
##  $ longitude                     : num  -74 -74 -73.9 -74 -73.9 ...
##  $ room_type                     : chr  "Private room" "Entire home/apt" "Private room" "Entire home/
##  $ price                         : int  149 225 150 89 80 200 60 79 79 150 ...
##  $ minimum_nights                : int  1 1 3 1 10 3 45 2 2 1 ...
##  $ number_of_reviews             : int  9 45 0 270 9 74 49 430 118 160 ...
##  $ last_review                   : chr  "10/19/2018" "5/21/2019" "" "7/5/2019" ...
##  $ reviews_per_month             : num  0.21 0.38 NA 4.64 0.1 0.59 0.4 3.47 0.99 1.33 ...
##  $ calculated_host_listings_count: int  6 2 1 1 1 1 1 1 1 4 ...
##  $ availability_365              : int  365 355 365 194 0 129 0 220 0 188 ...
```

```
summary(nyc)
```

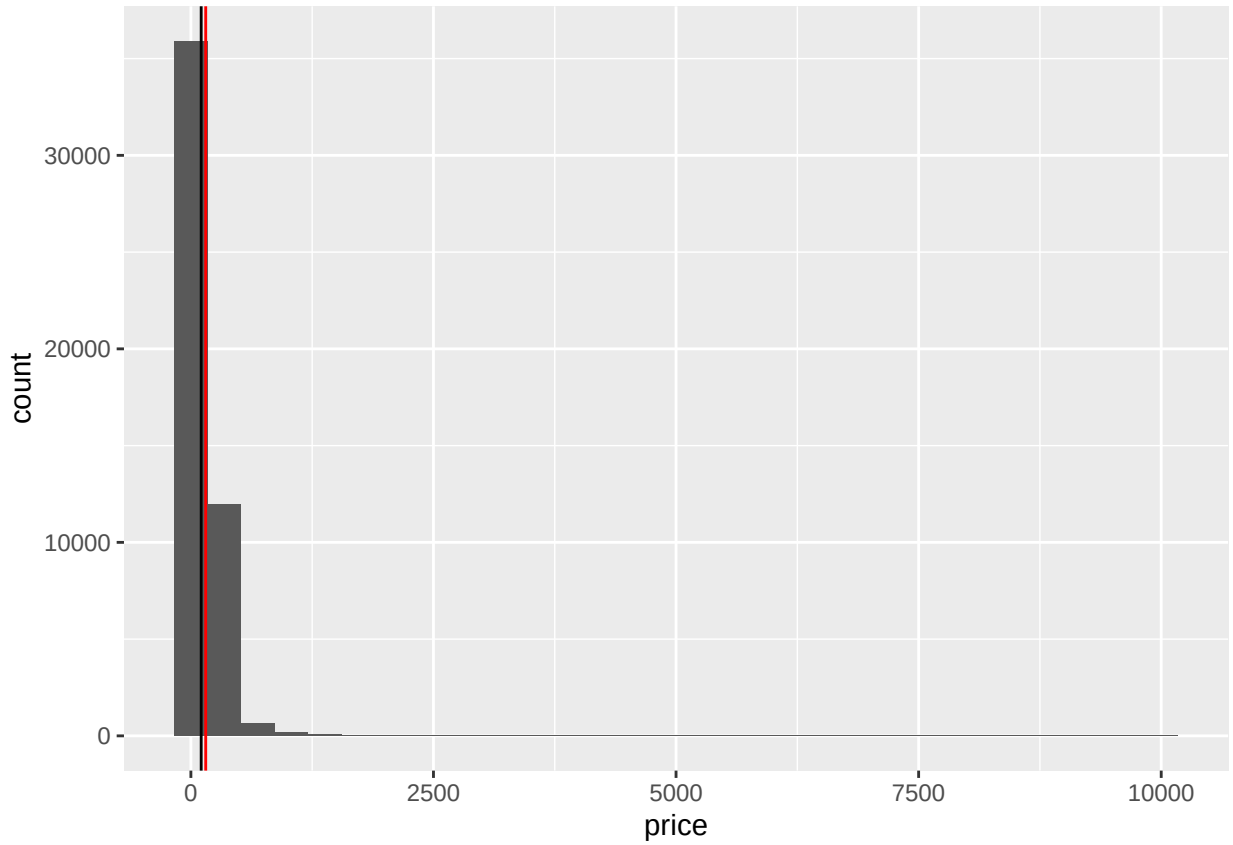
```
##        id                 name              host_id           host_name
##  Min.   :    2539   Length:48895       Min.   :     2438   Length:48895
##  1st Qu.: 9471945   Class :character   1st Qu.:  7822033   Class :character
##  Median :19677284   Mode  :character   Median : 30793816   Mode  :character
##  Mean   :19017143                      Mean   : 67620011
##  3rd Qu.:29152178                      3rd Qu.:107434423
##  Max.   :36487245                      Max.   :274321313
##
##  neighbourhood_group neighbourhood         latitude       longitude
##  Length:48895        Length:48895       Min.   :40.50   Min.   :-74.24
##  Class :character    Class :character   1st Qu.:40.69   1st Qu.:-73.98
##  Mode  :character    Mode  :character   Median :40.72   Median :-73.96
```

```
##                                      Mean   :40.73   Mean   :-73.95
##                                      3rd Qu.:40.76   3rd Qu.:-73.94
##                                      Max.   :40.91   Max.   :-73.71
##
##   room_type             price         minimum_nights    number_of_reviews
##  Length:48895       Min.   :    0.0   Min.   :   1.00   Min.   :  0.00
##  Class :character   1st Qu.:   69.0   1st Qu.:   1.00   1st Qu.:  1.00
##  Mode  :character   Median :  106.0   Median :   3.00   Median :  5.00
##                     Mean   :  152.7   Mean   :   7.03   Mean   : 23.27
##                     3rd Qu.:  175.0   3rd Qu.:   5.00   3rd Qu.: 24.00
##                     Max.   :10000.0   Max.   :1250.00   Max.   :629.00
##
##  last_review        reviews_per_month calculated_host_listings_count
##  Length:48895       Min.   : 0.010    Min.   :  1.000
##  Class :character   1st Qu.: 0.190    1st Qu.:  1.000
##  Mode  :character   Median : 0.720    Median :  1.000
##                     Mean   : 1.373    Mean   :  7.144
##                     3rd Qu.: 2.020    3rd Qu.:  2.000
##                     Max.   :58.500    Max.   :327.000
##                     NA's   :10052
##  availability_365
##  Min.   :  0.0
##  1st Qu.:  0.0
##  Median : 45.0
##  Mean   :112.8
##  3rd Qu.:227.0
##  Max.   :365.0
##
```

1. Create a histogram of the price feature and add mean and median lines to the plot. Describe in words what you can identify. (1 point)

```
ggplot(data=nyc, aes(x=price)) + geom_histogram() + geom_vline(xintercept=median(nyc$price)) + geom_vli
```

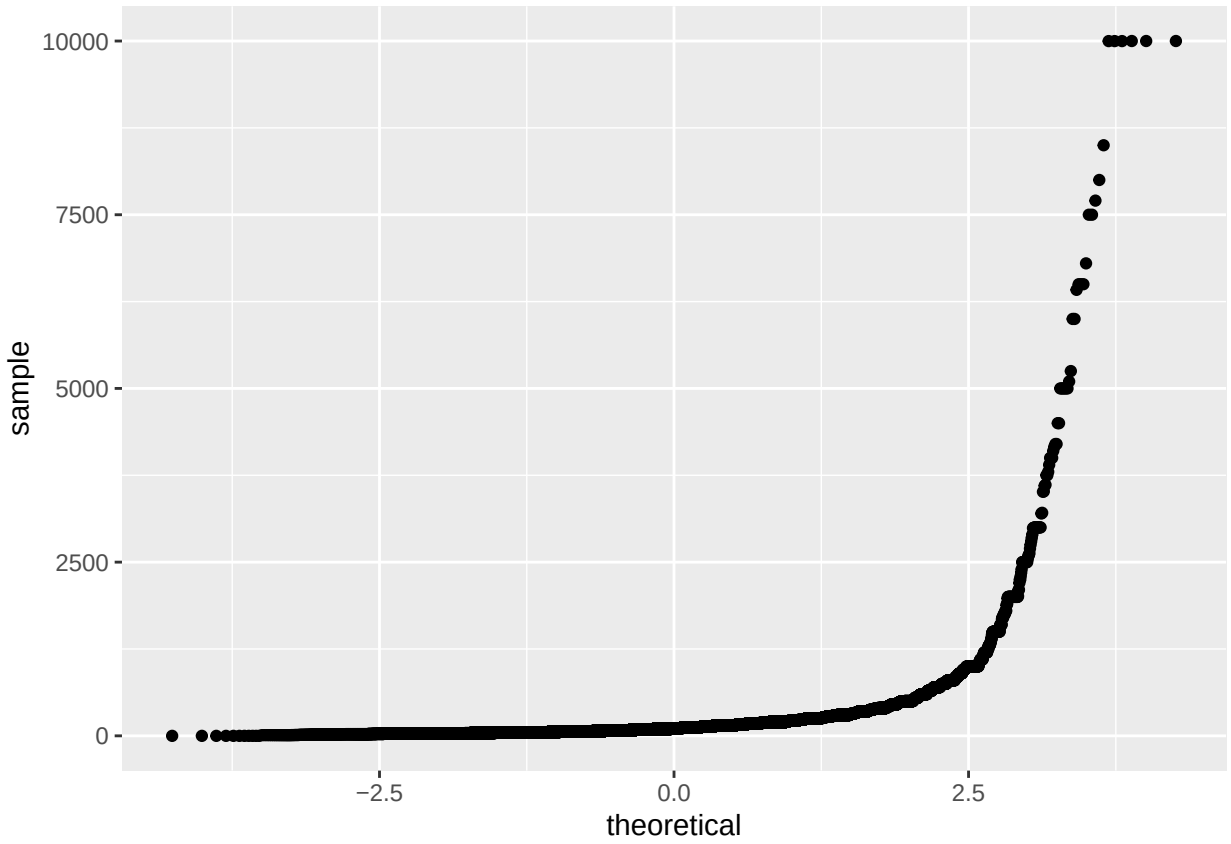
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



We can see that the mean (red) line is to the right of the median, which means that there are outliers to the right (since the mean is more sensitive to outliers) and that the distribution is skewed to the right.

2. Create the Q-Q plot for the price. Describe in words what assumptions you can make. (1 point)

```
ggplot(data=nyc, aes(sample=price)) + geom_qq()
```



Since the points aren't following a line, we can say that this distribution is not normal without testing for normality.

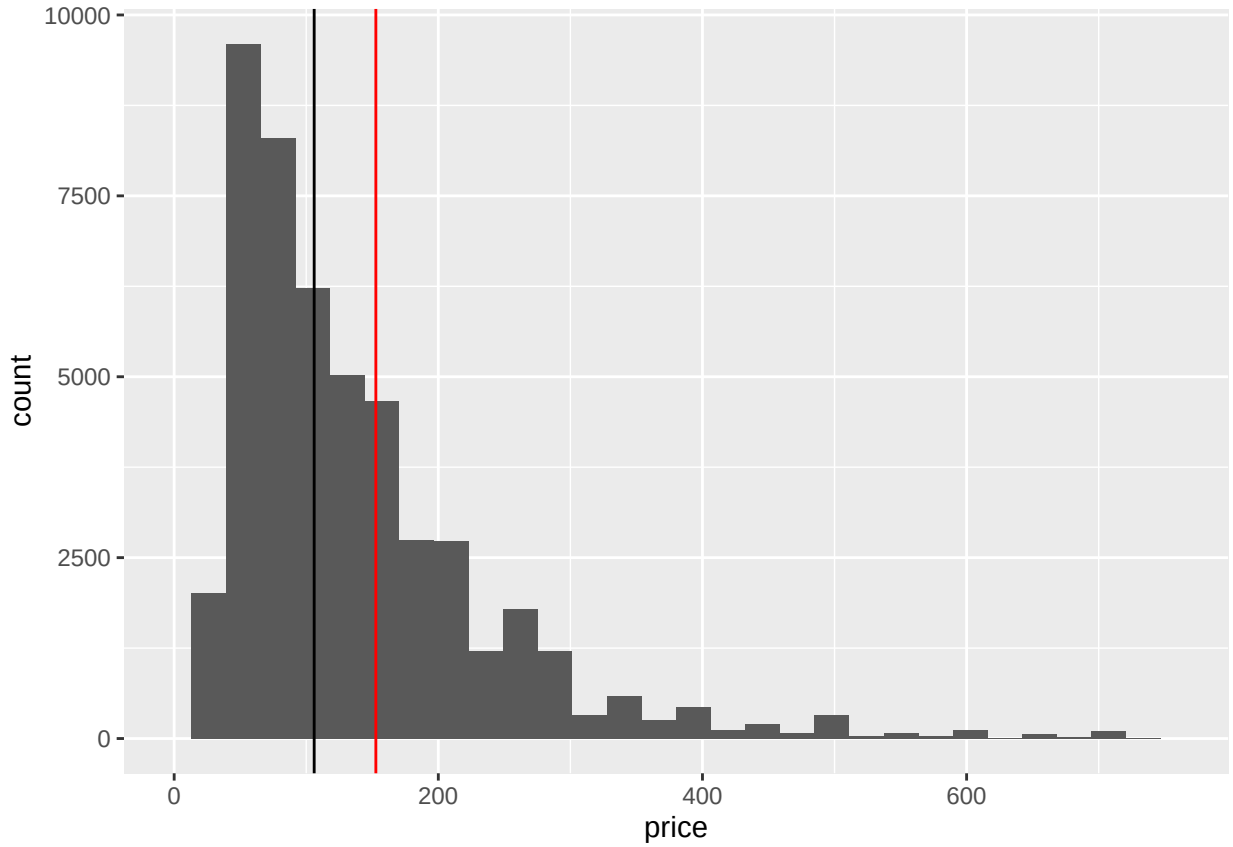
3. What do you think? Are there any changes that will make the plot more understandable? If yes, try to make that changes and continue using that plot for future tasks. (Hint play with the x-axis, try to change the limits) (1 point)

```
ggplot(data=nyc, aes(x=price)) + geom_histogram() + geom_vline(xintercept=median(nyc$price)) + geom_vli
```

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

```
## Warning: Removed 506 rows containing non-finite values (stat_bin).
```

```
## Warning: Removed 2 rows containing missing values (geom_bar).
```



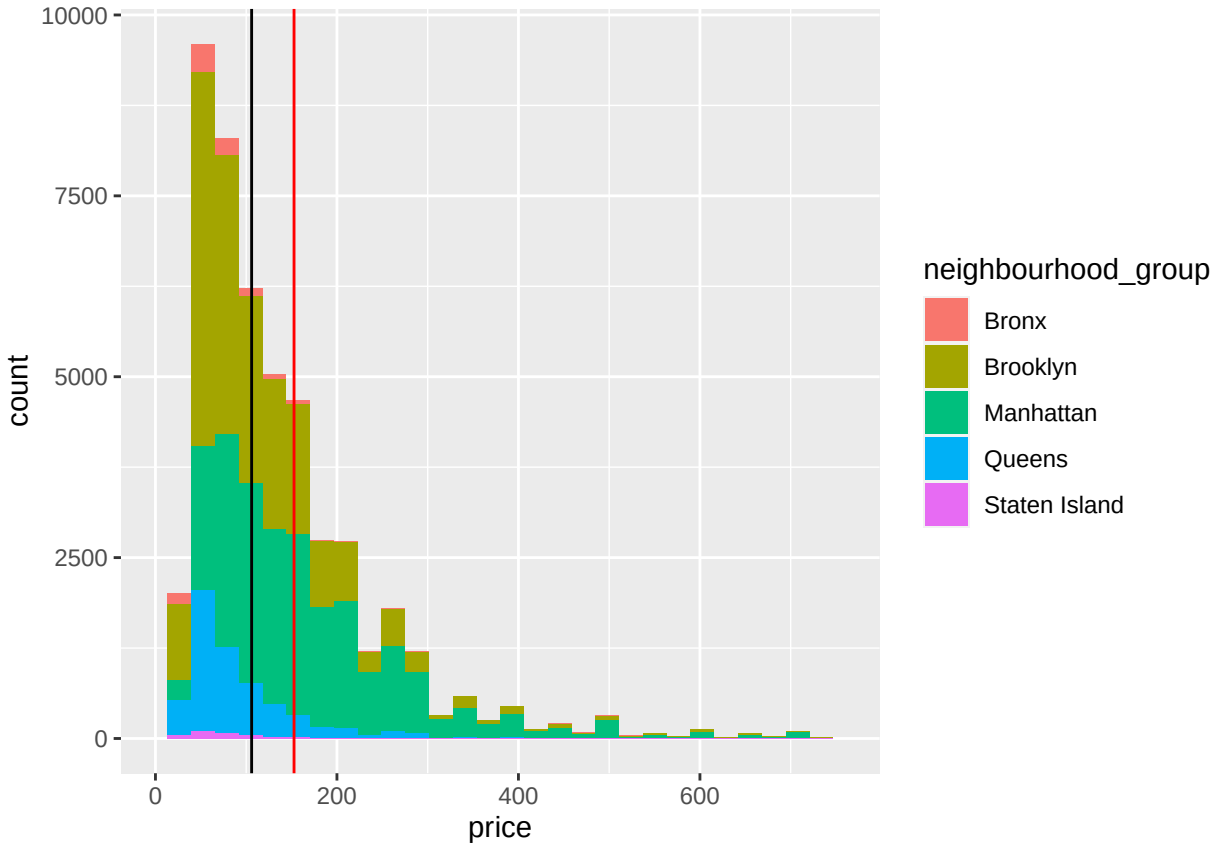
4. Create a stacked histogram and fill it with the categorical variable neighbourhood_group. Describe in words what you can identify. (1 point)

```
ggplot(data=nyc, aes(x=price, fill=neighbourhood_group)) + geom_histogram() + geom_vline(xintercept=medi
```

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

```
## Warning: Removed 506 rows containing non-finite values (stat_bin).
```

```
## Warning: Removed 10 rows containing missing values (geom_bar).
```



Queens has a smaller quantity of homes and has less outliers (homes priced > 400). Manhattan has more variance as it is more fat-tailed. There are a very small number of homes in Bronx compared to other neighborhoods, but Staten Island has the least.

5. (1 point)

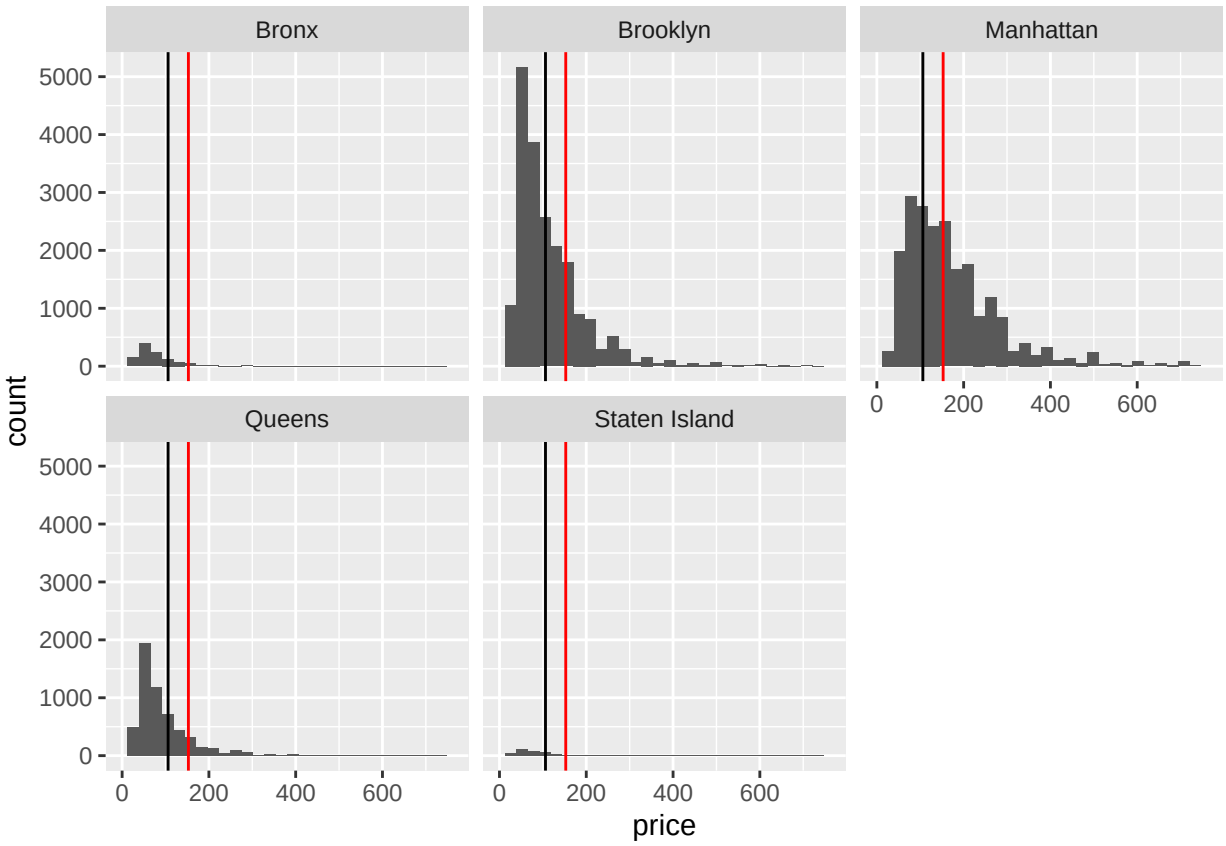
- 5.1.Use faceting to create a histogram for each neighborhood and remove the legend from the plot.

```
ggplot(data=nyc, aes(x=price)) + geom_histogram() + geom_vline(xintercept=median(nyc$price)) + geom_vli
```

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

```
## Warning: Removed 506 rows containing non-finite values (stat_bin).
```

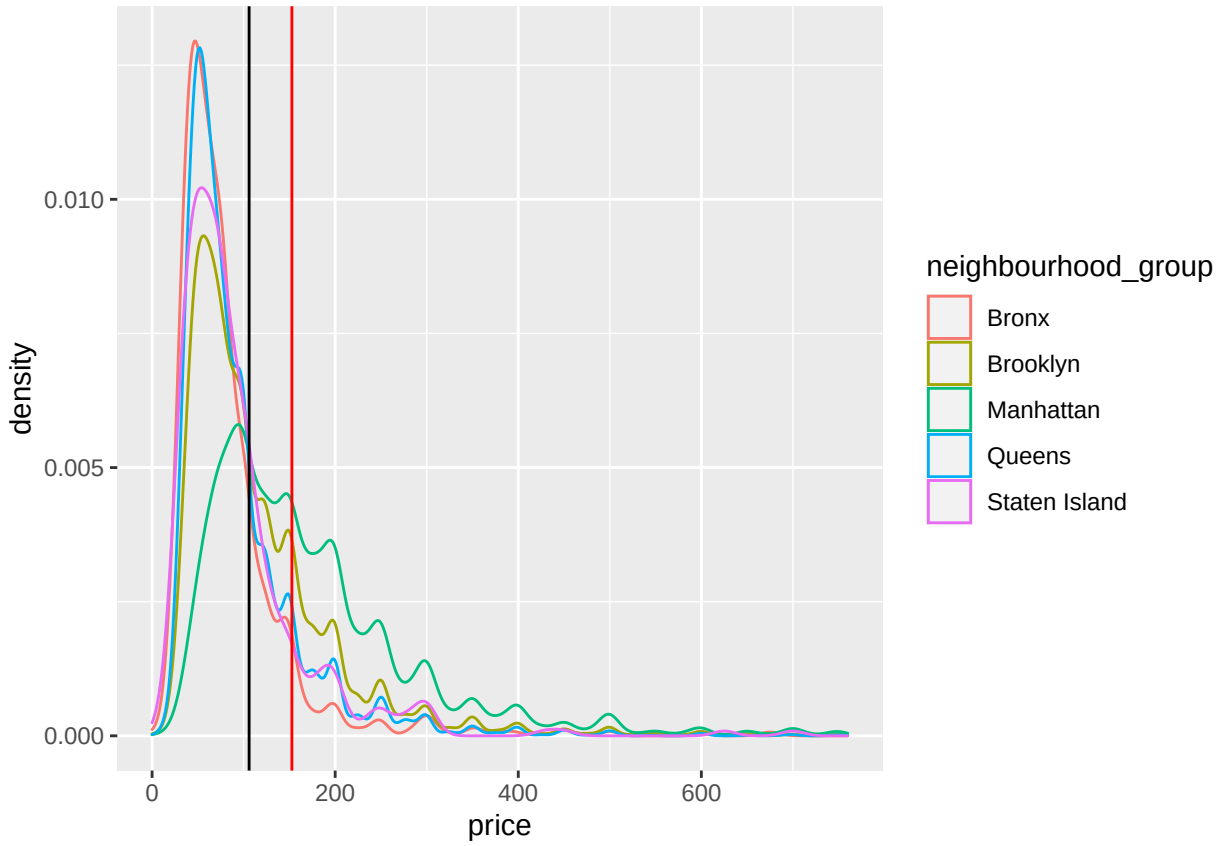
```
## Warning: Removed 10 rows containing missing values (geom_bar).
```



- 5.2.Why is it easier to compare distributions with density plots? Create the density plot of the price and fill it with the category of neighbourhood_group. Please describe in words.

```
ggplot(data=nyc, aes(x=price, color=neighbourhood_group)) + geom_density() + geom_vline(xintercept=media
```

```
## Warning: Removed 506 rows containing non-finite values (stat_density).
```



It allows us to see the distributions normalized by their counts, so if Staten Island was barely visible before, we now have a clear idea of how the data is distributed. I didn't fill, instead I colored the lines because with fill there was overplotting.

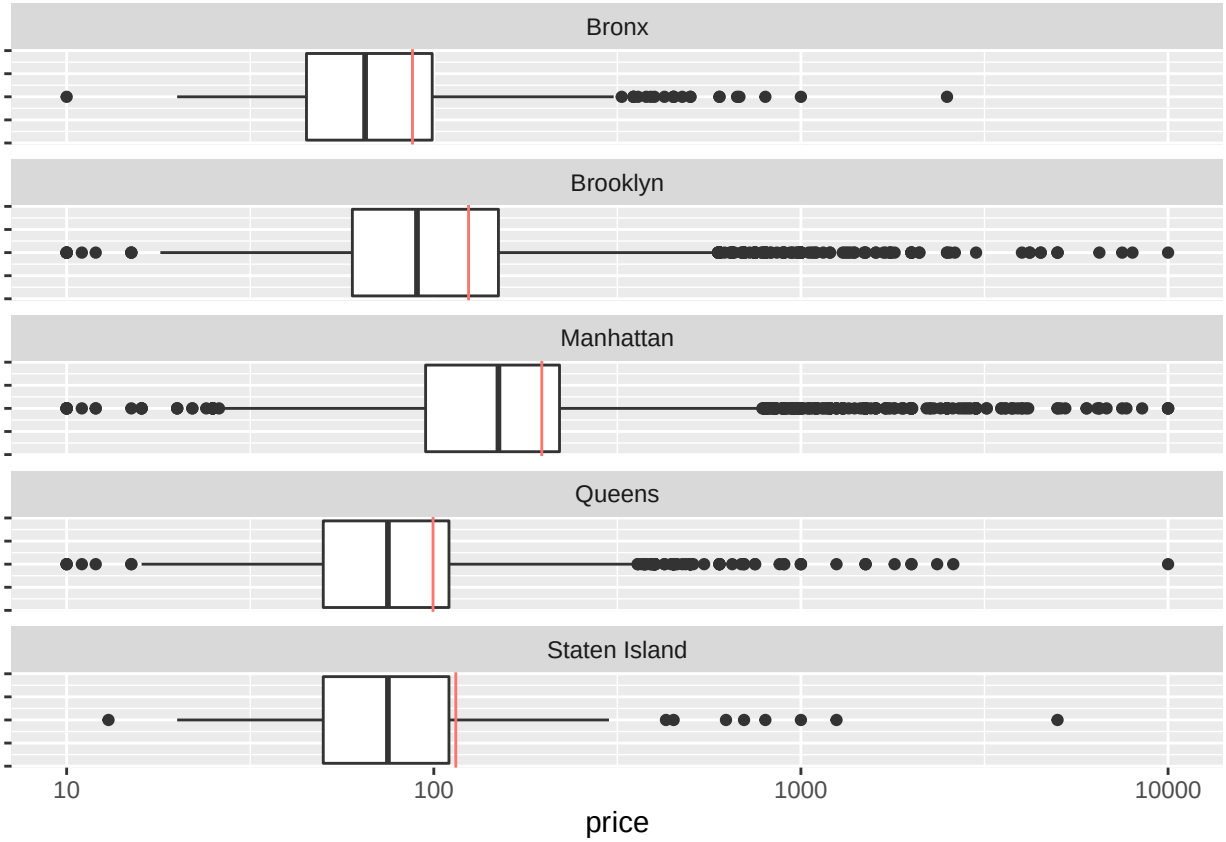
6. (1 point)

- 6.1.Create a boxplot for the price for each neighborhood and add the mean for each group with the red color. Are there any ways to compare the observations of each group? If yes, please apply it. Please describe it in words.

```
nyc %>% group_by(neighbourhood_group) %>%
  mutate(avg_p = mean(price)) %>%
  ggplot(aes(x=price)) + geom_boxplot() +
  theme(axis.text.y = element_blank()) +
  scale_x_continuous(trans = "log10") +
  facet_wrap(~ neighbourhood_group, nrow=5) +
  geom_vline(aes(xintercept=avg_p, color="red"), show.legend = FALSE)
```

```
## Warning: Transformation introduced infinite values in continuous x-axis
```

```
## Warning: Removed 11 rows containing non-finite values (stat_boxplot).
```



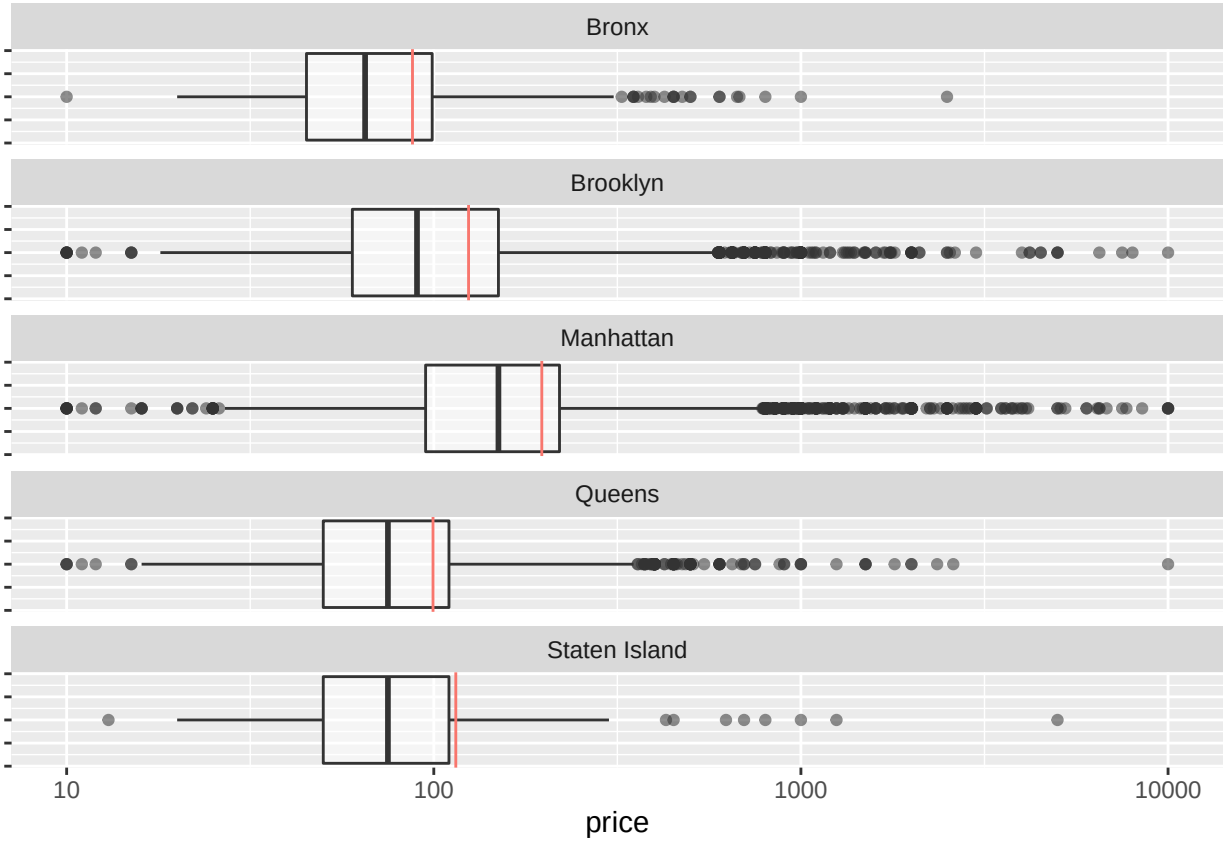
Yes, here we scale the x axis logarithmically to visualize the distribution better, then we can compare each distribution to each other since we made them share the x axis. Now we can clearly see that Manhatten is generally the more expensive neighborhood.

- 6.2.Do we have any outliers? If yes, please use the method to overcome overplotting.

```
nyc %>% group_by(neighbourhood_group) %>%
  mutate(avg_p = mean(price)) %>%
  ggplot(aes(x=price, alpha=0.01), show.legend = FALSE) +
  geom_boxplot(show.legend = FALSE) +
  theme(axis.text.y = element_blank()) +
  scale_x_continuous(trans = "log10") +
  facet_wrap(~ neighbourhood_group, nrow=5) +
  geom_vline(aes(xintercept=avg_p, color="red"), show.legend = FALSE)
```

```
## Warning: Transformation introduced infinite values in continuous x-axis
```

```
## Warning: Removed 11 rows containing non-finite values (stat_boxplot).
```



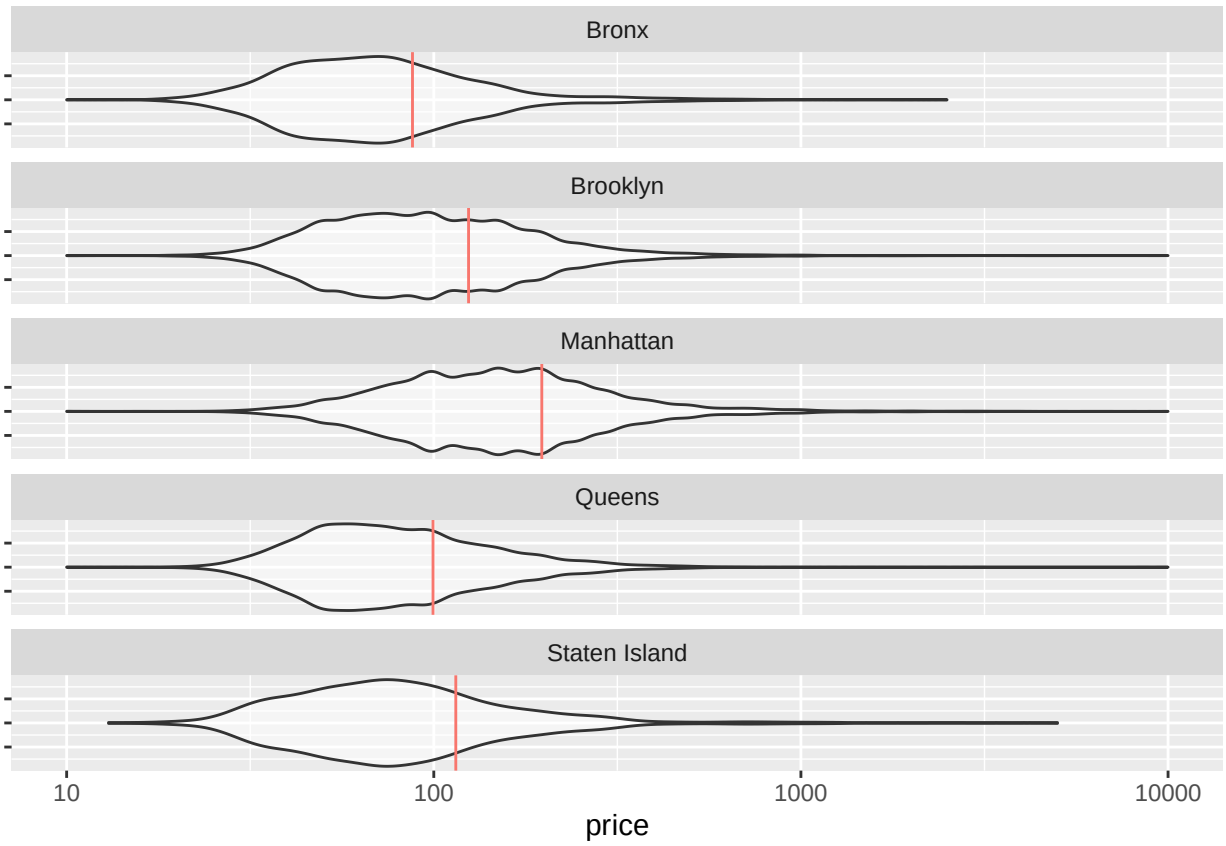
Yes, we overcame this issue by lowering opacity to 0.01

7. Create the violin plot for the price for each neighborhood. (1 point)

```
nyc %>% group_by(neighbourhood_group) %>%
  mutate(avg_p = mean(price)) %>%
  ggplot(aes(x=price, y=0, alpha=0.01), show.legend = FALSE) +
  geom_violin(show.legend = FALSE) +
  theme(axis.text.y = element_blank(), axis.title.y = element_blank()) +
  scale_x_continuous(trans = "log10") +
  facet_wrap(~ neighbourhood_group, nrow=5) +
  geom_vline(aes(xintercept=avg_p, color="red"), show.legend = FALSE)
```

```
## Warning: Transformation introduced infinite values in continuous x-axis
```

```
## Warning: Removed 11 rows containing non-finite values (stat_ydensity).
```



Comparing "Index of Economic Freedom" in different regions using Countries.csv data Source: The Heritage Foundation and The Wall Street Journal

Load the dataset Countries.csv. Check the structure and summary of the data.

```
ef <- read.csv("Countries.csv")
str(ef)
```

```
## 'data.frame':    61 obs. of  20 variables:
##  $ Country.Name           : chr  "Argentina" "Austria" "Bahamas" "Barbados" ...
##  $ Abbr                   : chr  "ARG" "AUT" "BHS" "BRB" ...
##  $ Region                 : chr  "America" "Europe" "America" "America" ...
##  $ Property.Rights        : num  32.4 86 45.3 55.5 50.9 83.3 43.5 25.7 55 62.5 ...
##  $ Judical.Effectiveness  : num  39.6 81.8 48.7 33 56.3 69.3 48.7 15.4 49.7 38.9 ...
##  $ Government.Integrity   : num  38.2 75.2 38.2 34.3 37.6 71.5 35 32.6 33.4 41.8 ...
##  $ Fiscal.Health          : num  56.4 79.7 42.3 0 92.8 66.3 60.5 81.4 22.8 86.4 ...
##  $ Business.Freedom       : num  57.3 76.9 68.5 69.6 71.3 82 62.7 58.9 61.3 66.7 ...
##  $ Labor.Freedom_2015     : num  46.1 67.6 71.5 67.7 74.6 61.1 53.6 35.8 52.3 68.3 ...
##  $ Labor.Freedom_2016     : num  52.3 65 88.1 79.3 87.7 ...
##  $ Monetary.Freedom       : num  50.9 83.4 77 83.7 60.4 84.9 79.6 66.4 67 83.3 ...
##  $ Trade.Freedom          : num  66.7 87 50.6 62.2 80.6 87 70.1 76 69.4 87 ...
##  $ Investment.Freedom     : int  50 90 50 75 30 85 50 5 50 70 ...
##  $ Financial.Freedom      : int  50 70 60 60 10 70 50 40 50 60 ...
##  $ GDP.Growth.Rate        : num  1.2 0.9 0.5 0.5 -3.9 1.4 1.5 4.8 -3.8 3 ...
##  $ GDP.per.Capita.PPP     : int  22554 47250 25167 16575 17654 43585 8373 6465 15615 19097 ...
##  $ Unemployment           : num  6.7 5.7 14.4 12.3 6.1 8.7 11.8 3.6 7.2 9.8 ...
##  $ Inflation.Perc         : num  26.5 0.8 1.9 0.5 13.5 0.6 -0.6 4.1 9 -1.1 ...
##  $ FDI.Inflow.Millions    : num  11655 3837 385 254 1584 ...
```

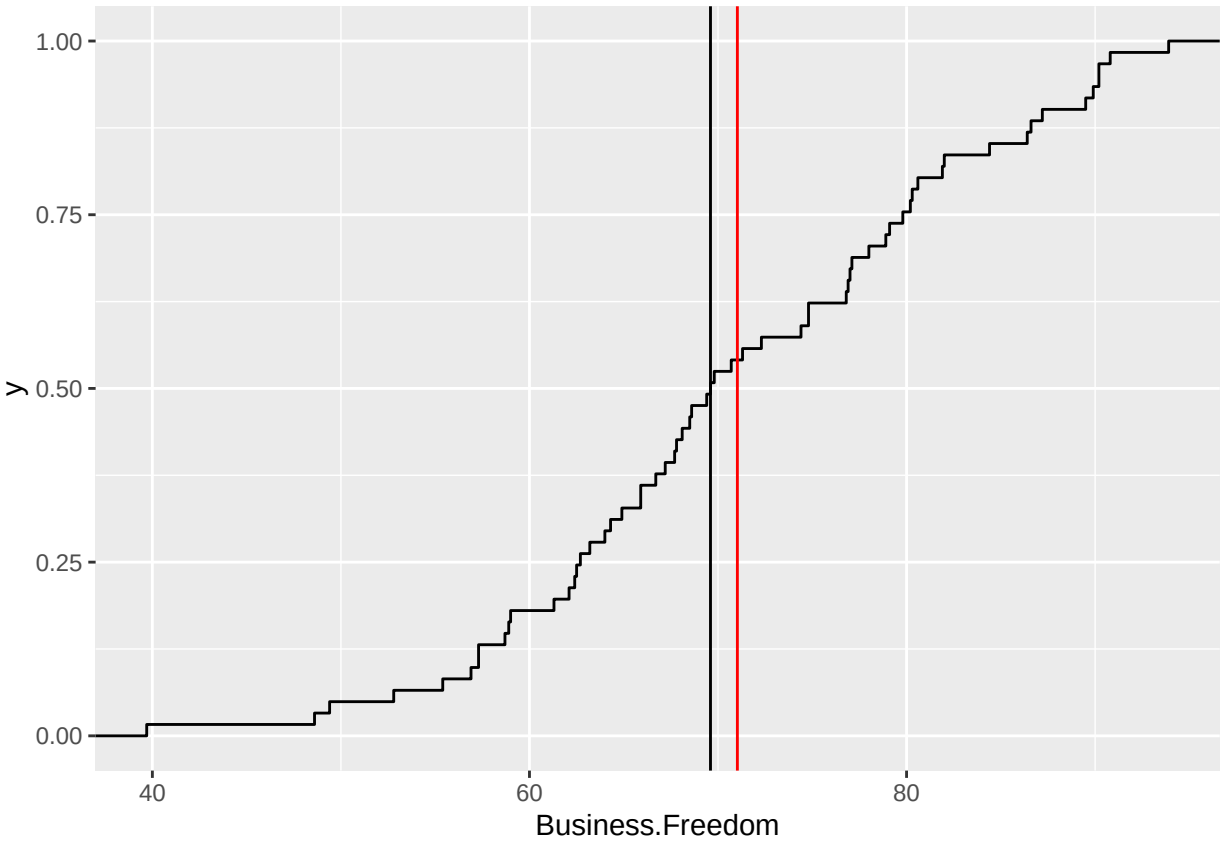
```
##  $ Public.Debt.Perc.of.GDP: num  56.5 86.2 65.7 103 59.9 ...
```

```
summary(ef)
```

```
##  Country.Name           Abbr               Region          Property.Rights
##  Length:61          Length:61          Length:61          Min.   : 6.80
##  Class :character   Class :character   Class :character   1st Qu.:47.60
##  Mode  :character   Mode  :character   Mode  :character   Median :61.30
##                                                           Mean   :62.27
##                                                           3rd Qu.:82.60
##                                                           Max.   :93.80
##  Judical.Effectiveness Government.Integrity Fiscal.Health  Business.Freedom
##  Min.   :10.30         Min.   :11.60        Min.   : 0.0   Min.   :39.70
##  1st Qu.:34.10         1st Qu.:33.90        1st Qu.:60.3   1st Qu.:62.70
##  Median :55.40         Median :41.80        Median :80.9   Median :69.60
##  Mean   :52.06         Mean   :51.78        Mean   :73.2   Mean   :71.03
##  3rd Qu.:69.30         3rd Qu.:70.50        3rd Qu.:93.4   3rd Qu.:79.80
##  Max.   :93.00         Max.   :90.00        Max.   :99.8   Max.   :93.90
##  Labor.Freedom_2015 Labor.Freedom_2016 Monetary.Freedom Trade.Freedom
##  Min.   :28.50      Min.   :21.76      Min.   :16.80    Min.   :50.60
##  1st Qu.:48.80      1st Qu.:52.81      1st Qu.:75.90    1st Qu.:77.80
##  Median :60.20      Median :62.55      Median :80.10    Median :87.00
##  Mean   :59.06      Mean   :62.54      Mean   :77.77    Mean   :81.44
##  3rd Qu.:70.90      3rd Qu.:72.96      3rd Qu.:85.00    3rd Qu.:87.00
##  Max.   :91.00      Max.   :97.34      Max.   :91.70    Max.   :90.00
##  Investment.Freedom Financial.Freedom GDP.Growth.Rate  GDP.per.Capita.PPP
##  Min.   : 0.00      Min.   :10.00     Min.   :-9.900   Min.   : 1750
##  1st Qu.:65.00      1st Qu.:50.00     1st Qu.: 1.000   1st Qu.:13847
##  Median :75.00      Median :60.00     Median : 2.100   Median :23460
##  Mean   :68.69      Mean   :58.03     Mean   : 1.834   Mean   :27051
##  3rd Qu.:80.00      3rd Qu.:70.00     3rd Qu.: 3.600   3rd Qu.:41120
##  Max.   :95.00      Max.   :90.00     Max.   : 7.800   Max.   :98987
##   Unemployment    Inflation.Perc    FDI.Inflow.Millions Public.Debt.Perc.of.GDP
##  Min.   : 2.700   Min.   : -1.100   Min.   : -4238.6    Min.   : 10.10
##  1st Qu.: 5.300   1st Qu.:  0.100   1st Qu.:   802.5    1st Qu.: 38.80
##  Median : 6.900   Median :  0.800   Median :  2221.5    Median : 49.40
##  Mean   : 7.564   Mean   :  5.449   Mean   : 18353.0    Mean   : 56.84
##  3rd Qu.: 9.800   3rd Qu.:  4.100   3rd Qu.: 12579.4    3rd Qu.: 73.70
##  Max.   :14.400   Max.   :121.700   Max.   :379894.0    Max.   :132.60
```

8. Create the ecDF of Business. Freedom. To make the plot more informative, add the mean and the median to the plot. What assumptions can you make? Please describe in words. (1 point)

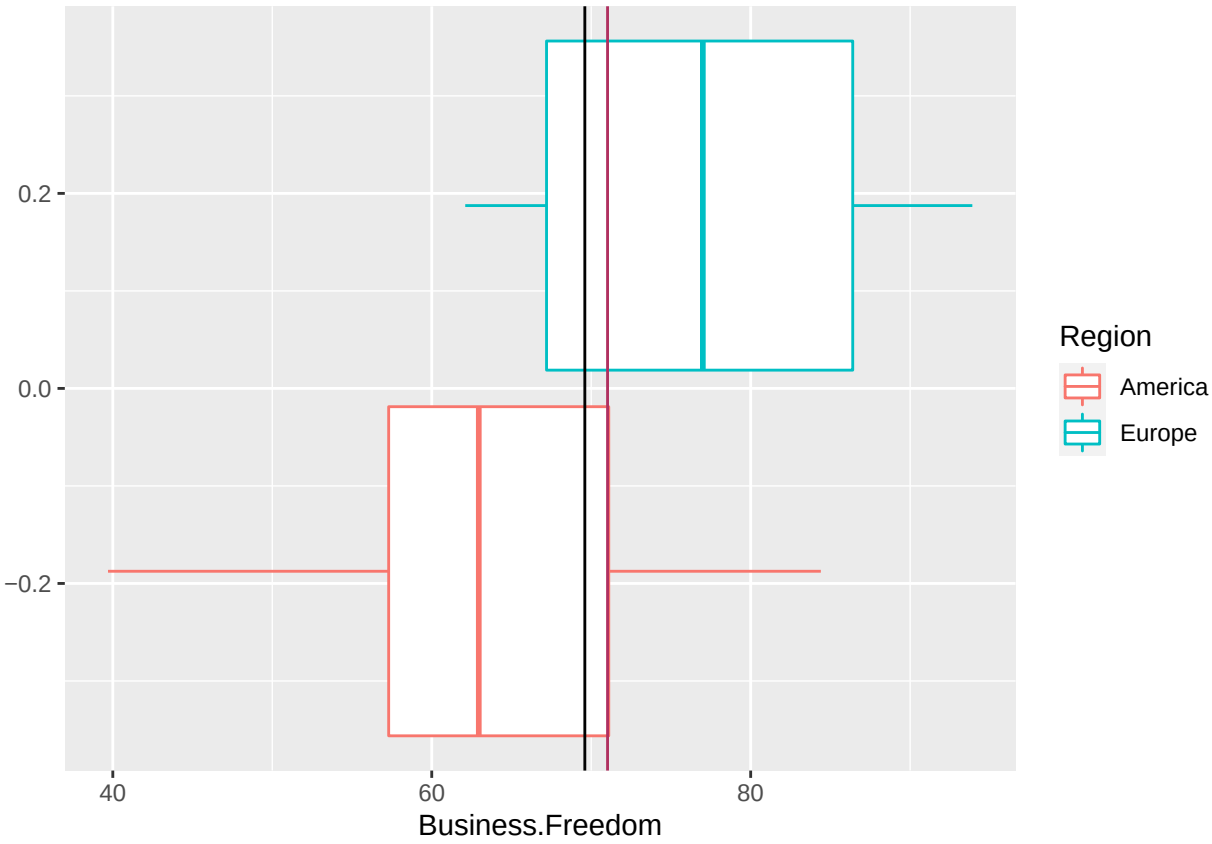
```
ef %>% ggplot(aes(x=Business.Freedom)) +
  stat_ecdf() +
  geom_vline(xintercept=mean(ef$Business.Freedom), color="red") +
  geom_vline(xintercept=median(ef$Business.Freedom))
```



It is slightly right-skewed, the lowest value is 40 and the highest is about 95

9. Is the Business Freedom Rate the same in Europe and America? (1 point)

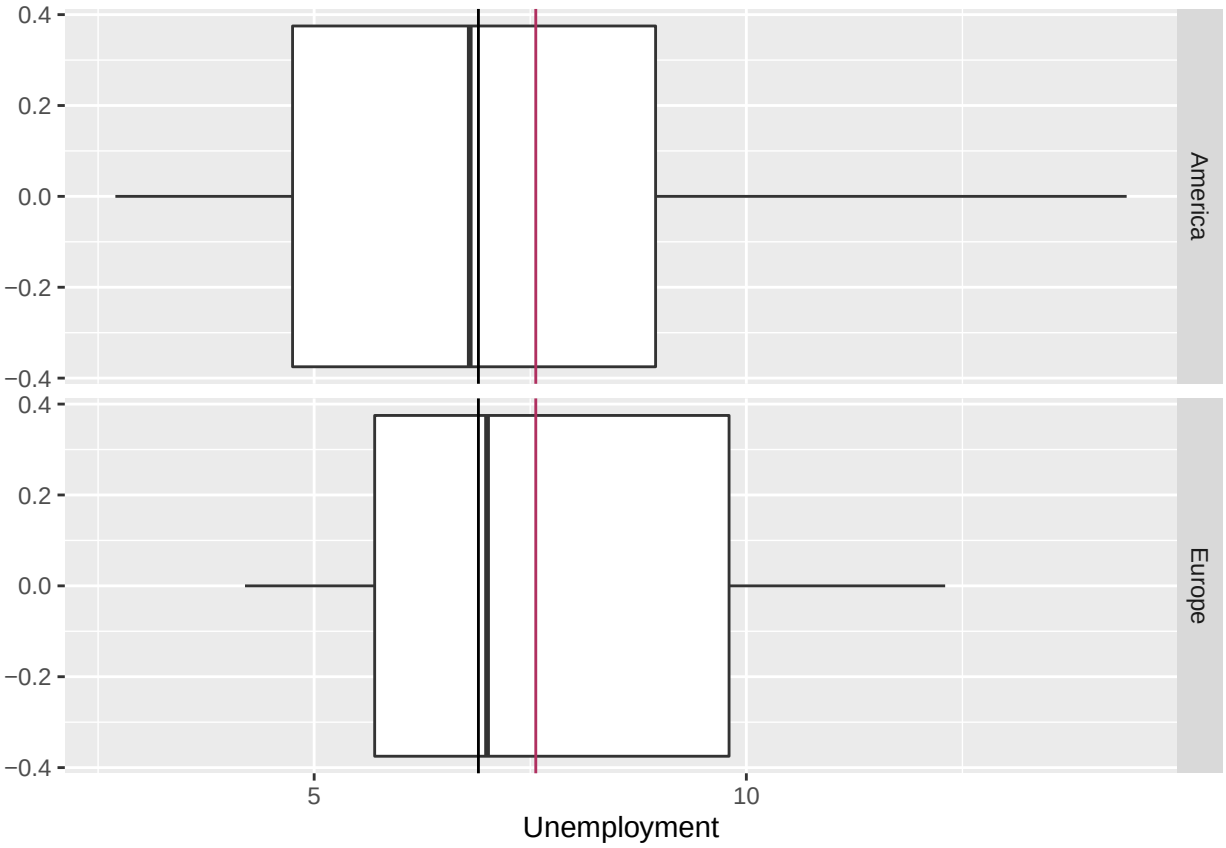
```
ef %>% ggplot(aes(x=Business.Freedom, color=Region)) +
  geom_boxplot() +
  geom_vline(xintercept=mean(ef$Business.Freedom), color="maroon") +
  geom_vline(xintercept=median(ef$Business.Freedom))
```



Business freedom in Europe is better than in America

10. Arrange various distribution plots in a grid using the Unemployment feature. (1 point)

```
ef %>% ggplot(aes(x=Unemployment)) +
  geom_boxplot() +
  geom_vline(xintercept=mean(ef$Unemployment), color="maroon") +
  geom_vline(xintercept=median(ef$Unemployment)) +
  facet_grid(Region ~ .)
```



To do more plots we would need more non-unique categorical variables