DataVizA Tutorial: Basic Visualisation: Solutions

Department of Econometrics and Business Statistics, Monash University

Tutorial 3

Concepts

1. In the RGB color model what color is represented by #FF0000?

Since RGB stands for Red Green Blue the color FF0000 is the maximum possible amount of red (FF=255) and no green and blue. That is FF0000 is pure red.

2. In the RGB color model what would be the hex code for pure blue?

This would be #0000FF.

3. Describe the role of bandwidth in kernel density estimation.

A kernel density estimated at a value x gives highest weight to values near x. The bandwidth essentially determines how close is meant by 'near'. Large values lead to smooth (in the limit flat) density estimates, small values lead to very bumpy estimates.

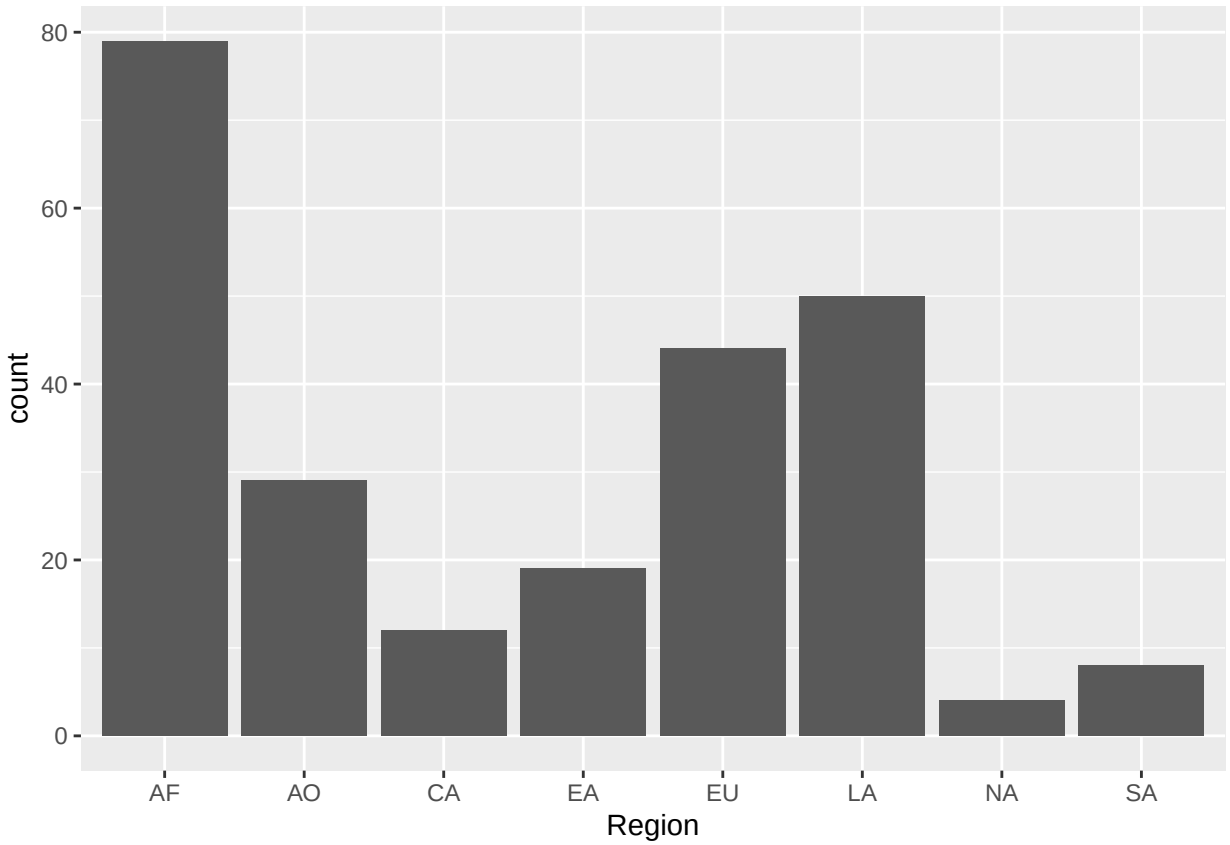
4. How are outliers defined in a boxplot?

An upper and lower fence are set to the the third quartile plus 1.5 times the interquartile range and the first quartile minus 1.5 times the interquartile range respectively. If the minimum value is greater than the first quartile minus 1.5 of the interquartile range, then the a line is drawn at the mimimum value rather than the lower fence. If the maximum value is less than the third quartile plus 1.5 of the interquartile range, then the upper fence is drawn at the maximum value rather than the upper fence. If there are points either above the upper fence, or below the lower fence these are considered to be outliers and drawn as a point.

Data Analysis

1. Using the Swiss Export Data, plot a bar chart of the region variable.

```
library(tidyverse)
SwissExp<-readRDS('SwissExport.rds')
ggplot(SwissExp,aes(x=Region))+geom_bar()
```

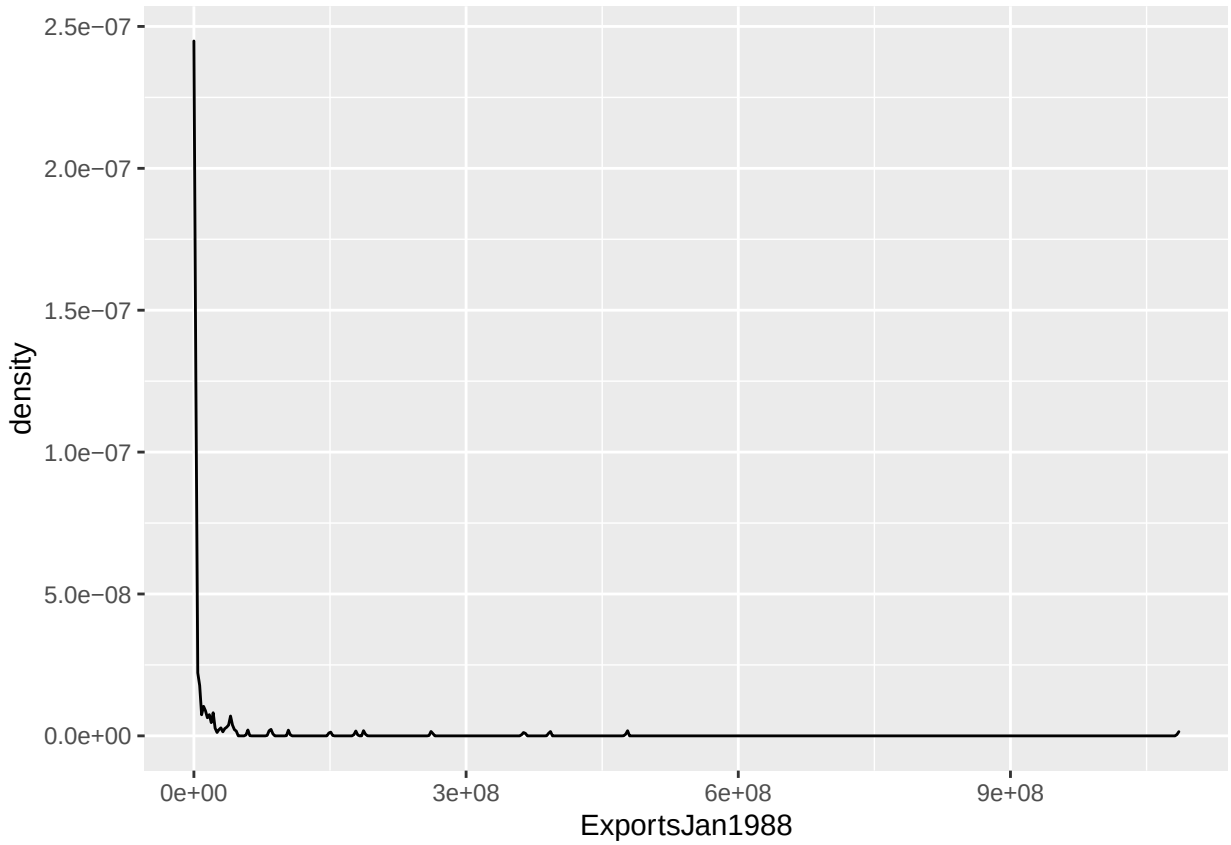


2. What does this plot tell you?

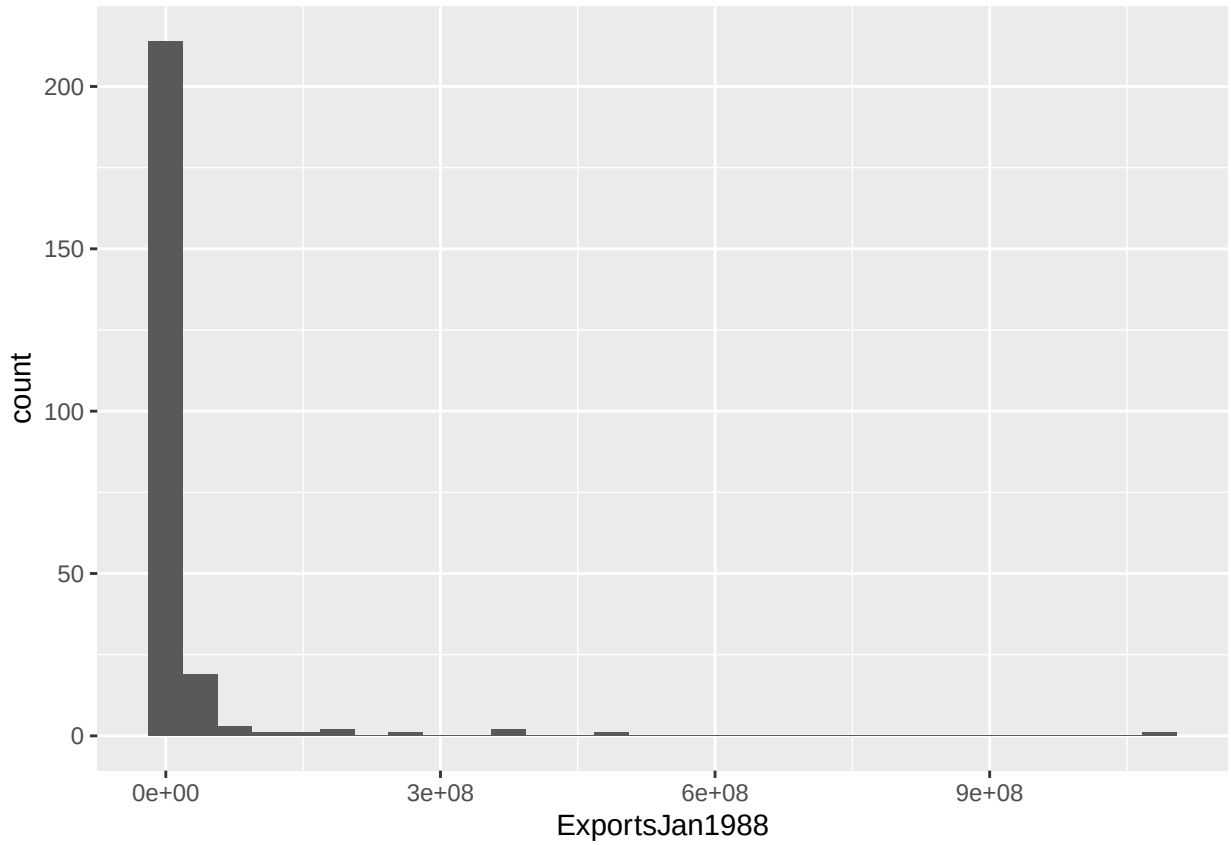
The plot merely states that most countries are in Africa and very few are in North America. This plot says nothing about Swiss Exports. In many cases this plot would not be worth including in an analysis.

3. Construct a density plot, rug plot and boxplot for Exports in January 1988.

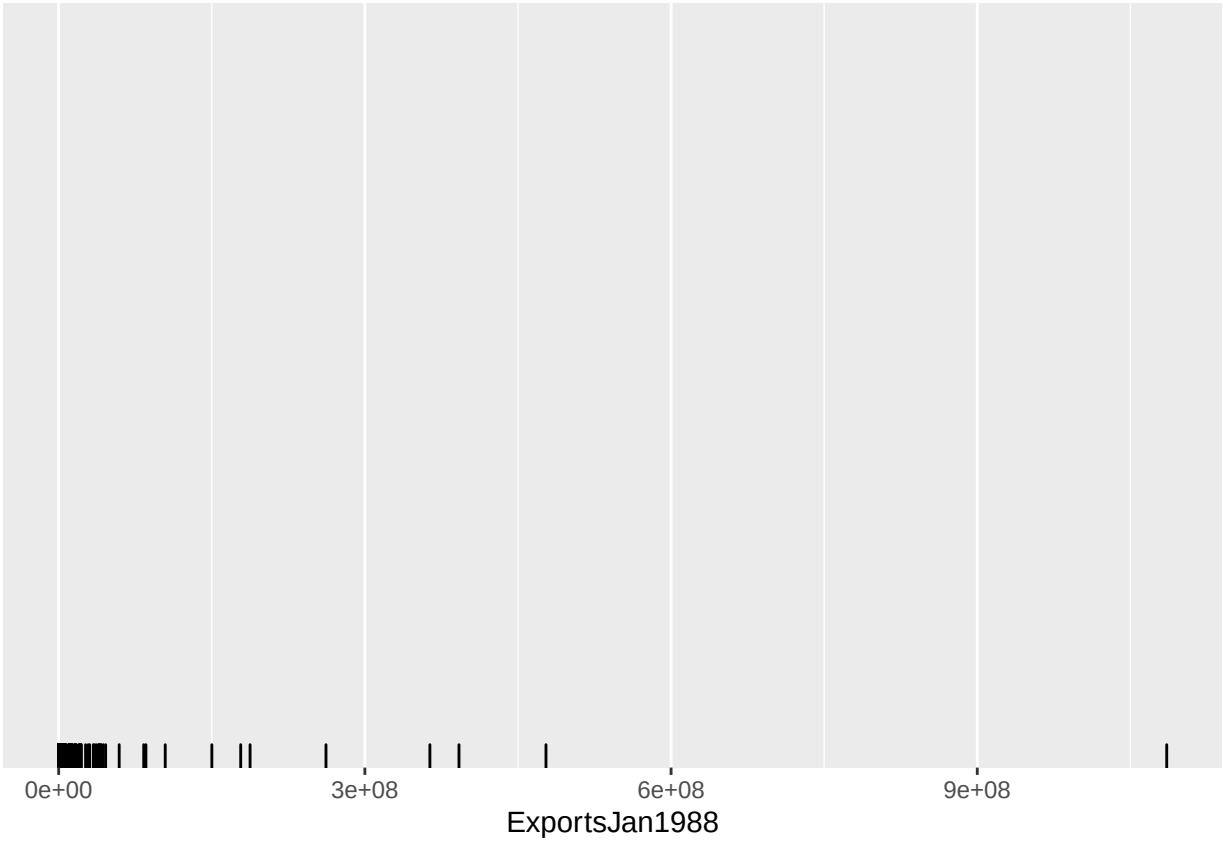
```
library(tidyverse)
SwissExp<-readRDS('SwissExport.rds')
ggplot(SwissExp,aes(x=ExportsJan1988))+geom_density()
```



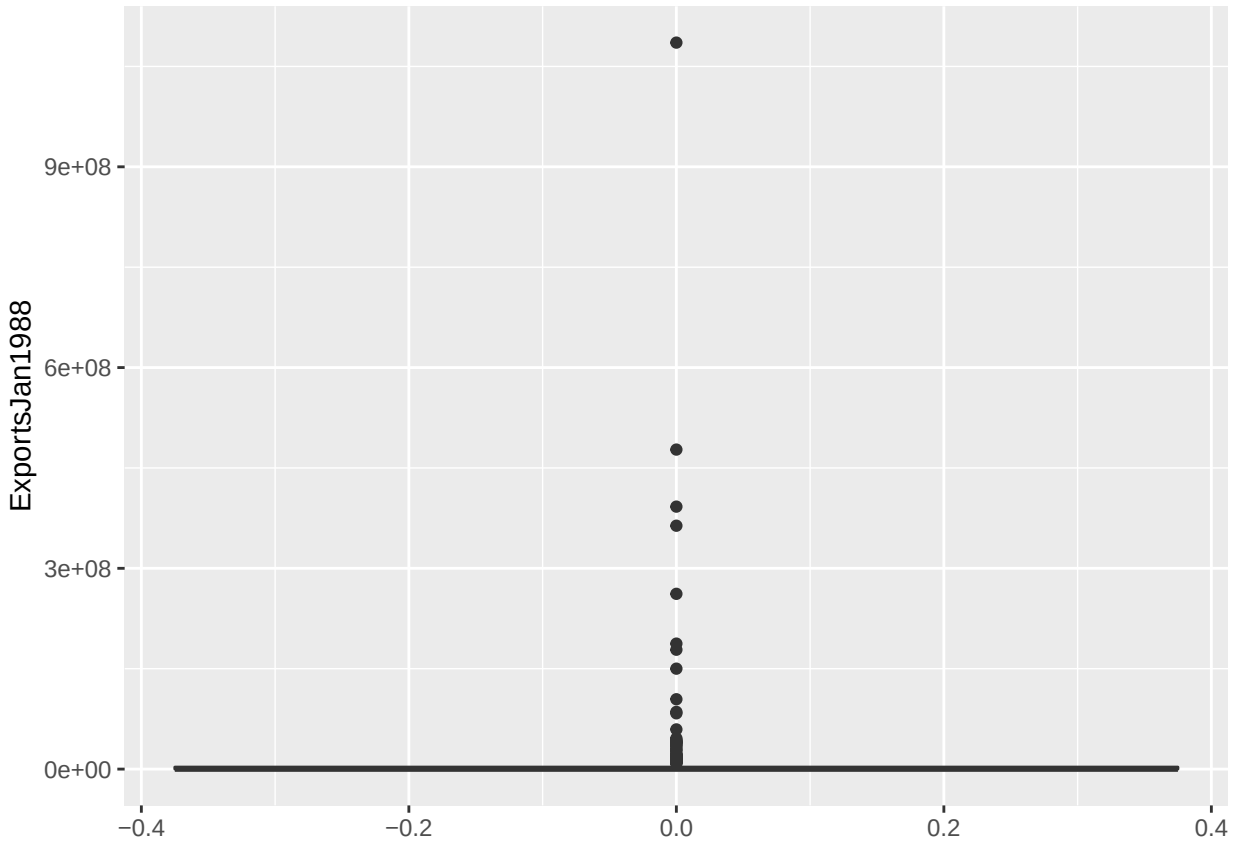

```
ggplot(SwissExp,aes(x=ExportsJan1988))+geom_histogram()
```




```
ggplot(SwissExp,aes(x=ExportsJan1988))+geom_rug()
```




```
ggplot(SwissExp,aes(y=ExportsJan1988))+geom_boxplot()
```



4. What do we see in these plots?

All plots show that the Switzerland had a very small level of exports to most countries. In the box plots, the quartiles and medians are all close to 0. There is one very large export destination which is Germany.

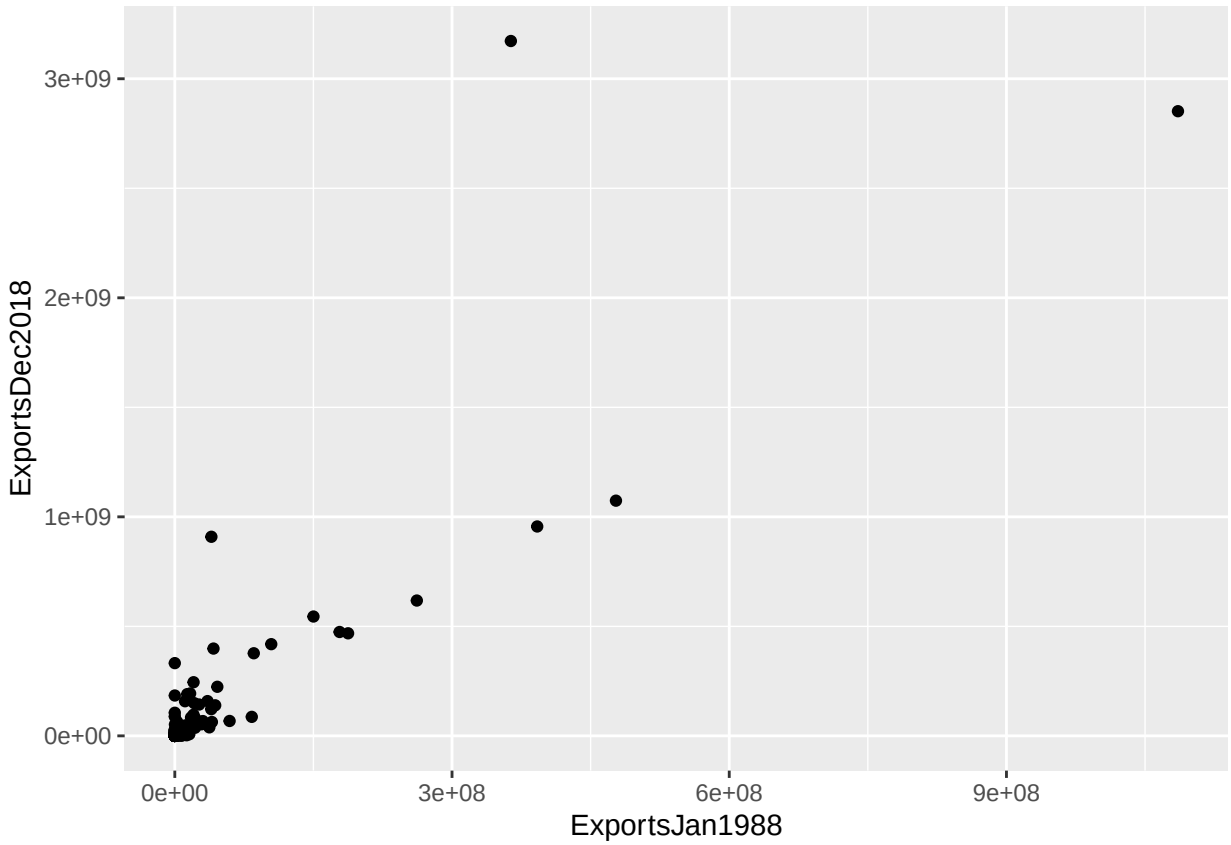
Log Scales

1. Why might we display data using a log scale?

Some data may be heavily skewed to the right. In such cases a log scale compresses together large values and stretches out small values.

2. Using the Swiss Exports data plot a scatterplot of exports in January 1988 against exports in December 2018.

```
SwissExp<-readRDS('SwissExport.rds')
library(tidyverse)
ggplot(SwissExp,
       aes(x=ExportsJan1988,
           y=ExportsDec2018))+
         geom_point()
```

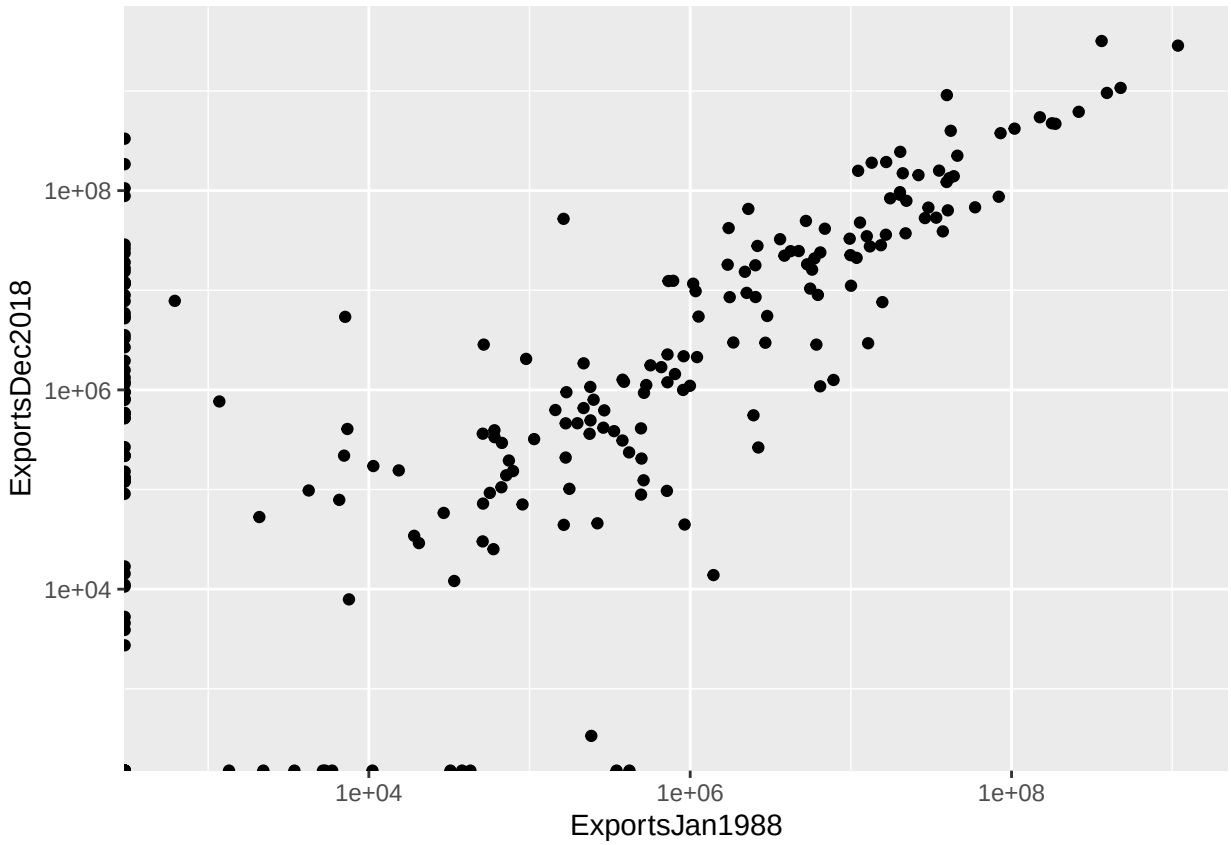


3. Do the same plot but using log scales.

```
SwissExp<-readRDS('SwissExport.rds')
library(tidyverse)
ggplot(SwissExp,
       aes(x=ExportsJan1988,
           y=ExportsDec2018))+
         geom_point()+
  scale_x_log10()+scale_y_log10()
```

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

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



4. Compare these two plots

On the log log scale there is less overplotting. Rather than have all countries in the bottom left hand corner the points are spread out over the whole plot.

5. How do you understand the warning message that occurs when the log log scale is used?

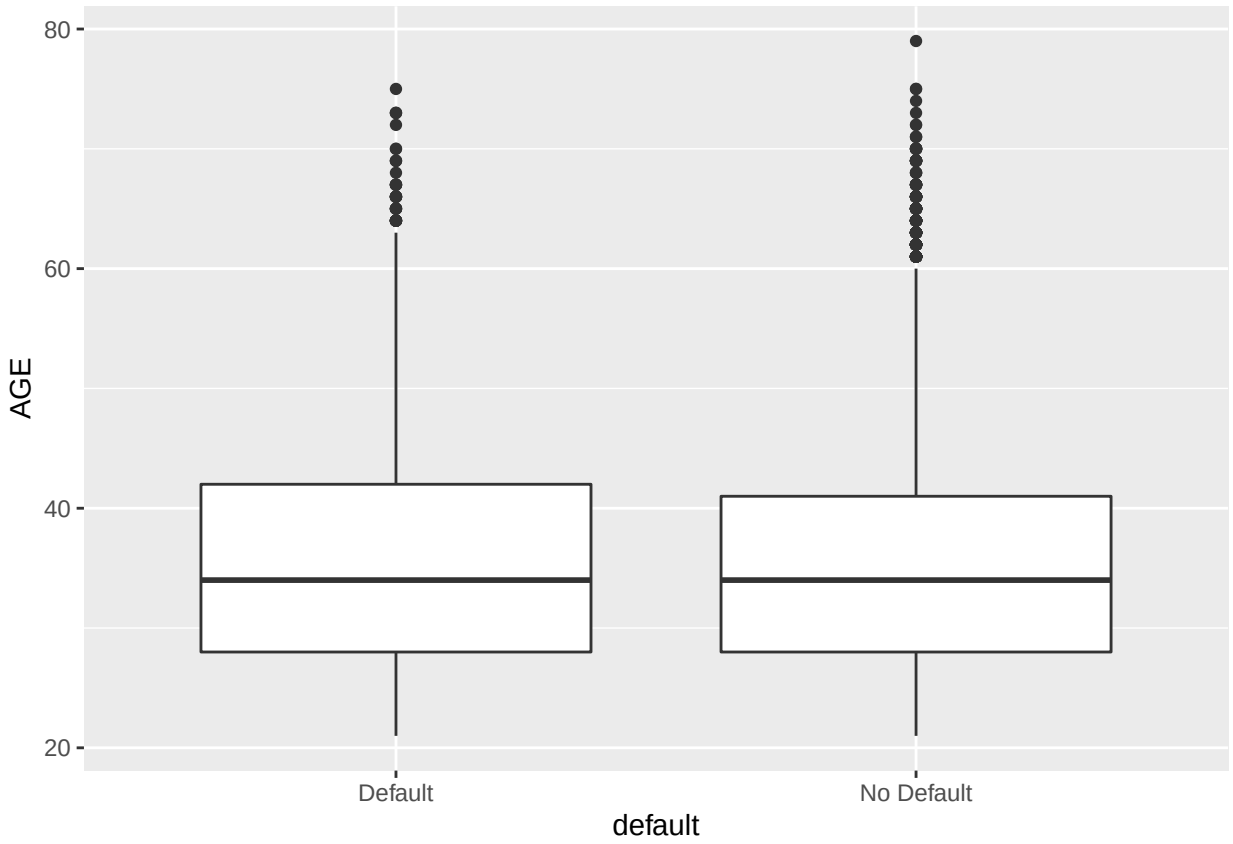
There are a number of countries that recieve no exports from Switzerland. These are values of 0. Mathematically, when taking the log of zero, the result is negative infinity. Note that ggplot has done something sensible and plotted these zero values along the margins of the plot. However, it may be better to either add a very small value to these zeroes or exclude them altogether.

Credit default data

The dataset *credit.rds* contains demographic information and repayment history for individuals who may have either defaulted on their credit card payments. The variable of interest is *default* which is equal to 1 for customers who fail to pay their credit card bill and 0 otherwise. More details on the dataset can be found here

1. Using box plots, explore whether the distribution of age is different for the default group and non-default group. Interpret your result.

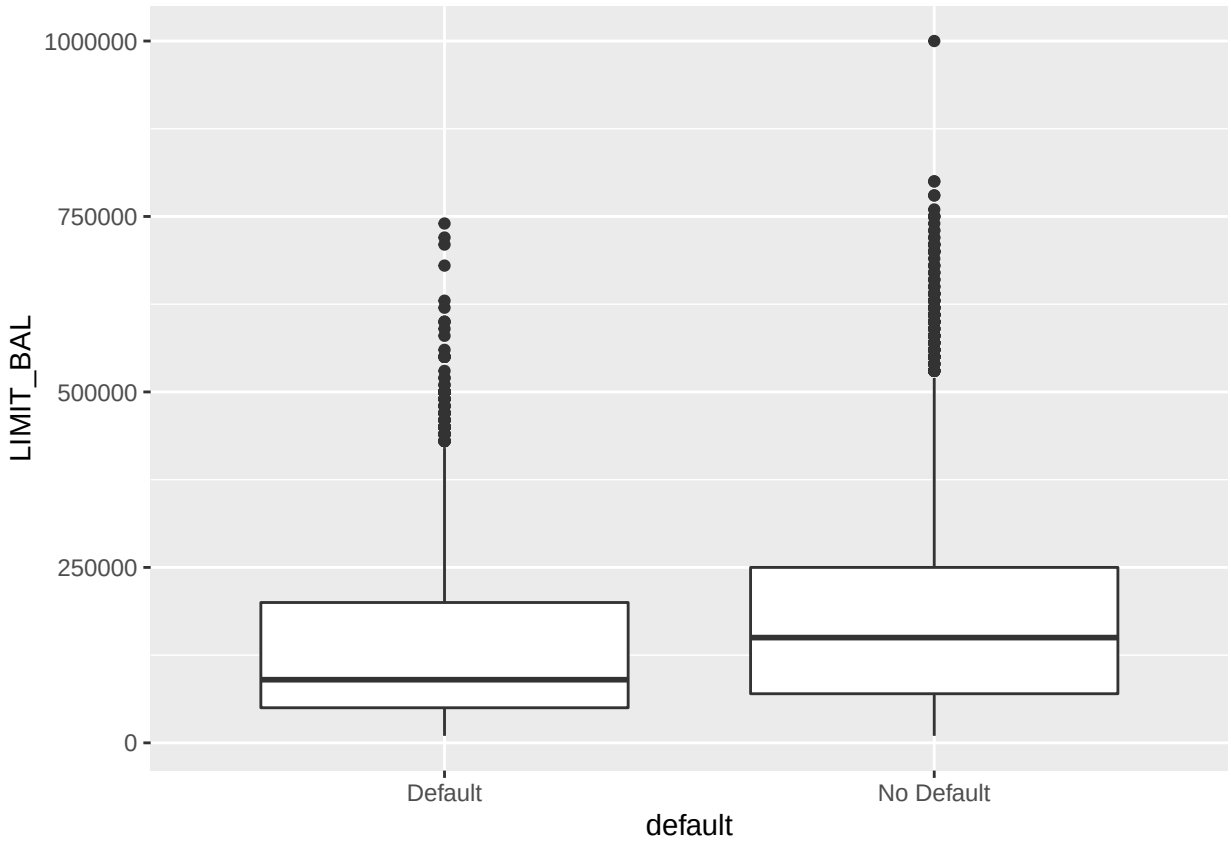
```
credit<-readRDS('Credit.rds')
ggplot(credit,
       aes(x=default,y=AGE))+
  geom_boxplot()
```




```
#The plot does not suggest differences between the age distribution
#of the default and non-default groups.  Age is not likely to be a
#useful feature in disciminating between these groups.
```

2. Using box plots, explore whether the distribution of the credit limit (LIMIT_BAL) is different for the default group and non-default group. Interpret your result.

```
credit<-readRDS('Credit.rds')
ggplot(credit,
       aes(x=default,y=LIMIT_BAL))+
  geom_boxplot()
```




```
#The plot suggests that distribution of the credit limit is higher
#for non-default groups.
```

3. Suppose you work for the credit card company. In a business meeting a colleague suggests that the credit card limit could be useful in predcting when a customer defaults.

The credit card limit could be useful in helping to explain default since a higher limit is associated with a lower probability of default.

4. Suppose you work for the credit card company. In a business meeting a colleague looks at the plot and suggests that low credit card limits are causing defaults. They suggest to raise the credit card limit of all customers. Is this a good idea?

This is probably a bad idea. Association does not imply causation. If anything the causation is likely to run the other way. Customers identified as being at risk of default for some other reason (unstable employment, low income) are only offered a low credit limit.