Coding task 21.01.2021: Comapring distributions

Contents

1 Example data & libraries

```
# install.packages("rattle.data")
# install.packages("ggridges")
library(rattle.data)
library(ggplot2)
library(ggpubr)

wine = wine
```

2 Different ways to visualise distributions

2.1 Plot the distribution of the wine color for the three different wine types

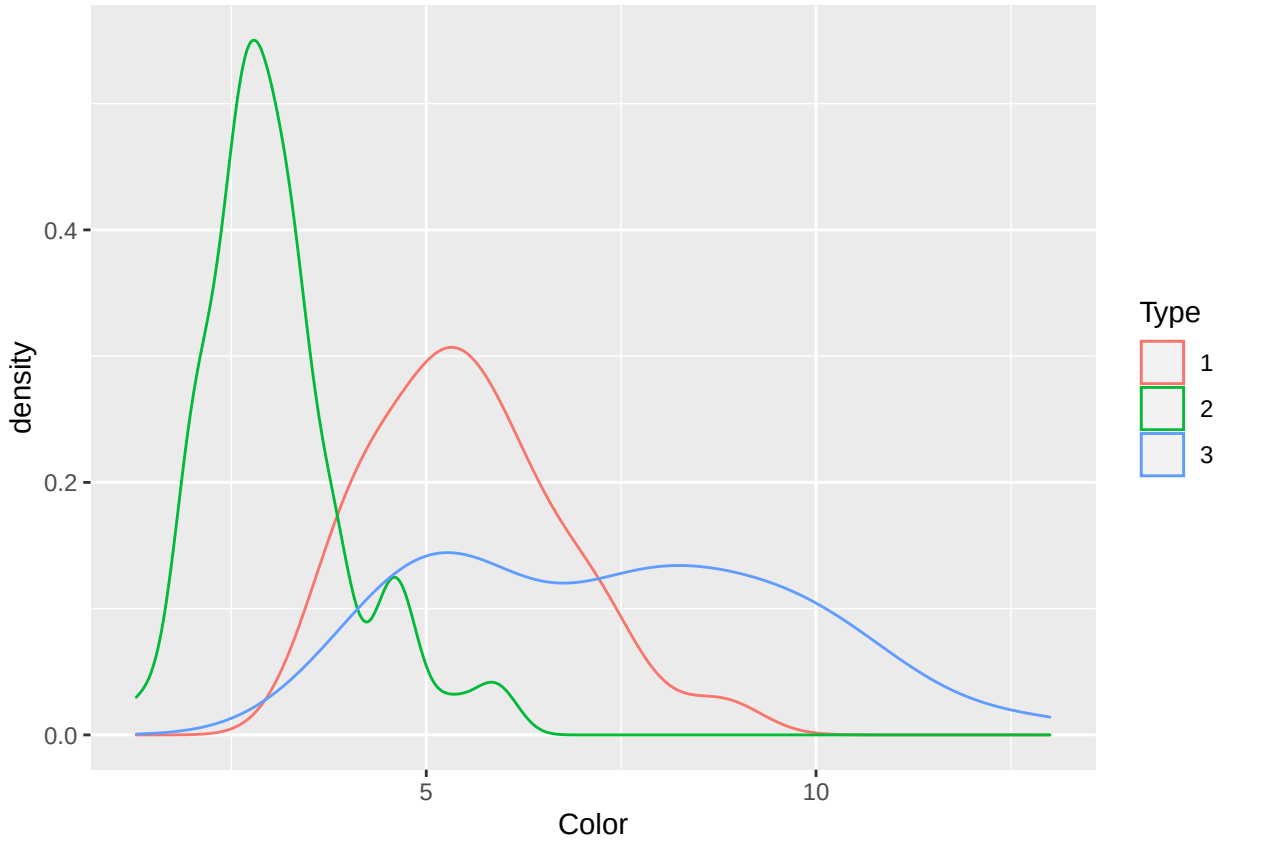
- Try to come up with different way to plot this
- Which is you favorite and why?

2.2 Here are some more, try them out

2.2.1 geom_density(), ggridges::geom_density_ridges(), geom_histogram()

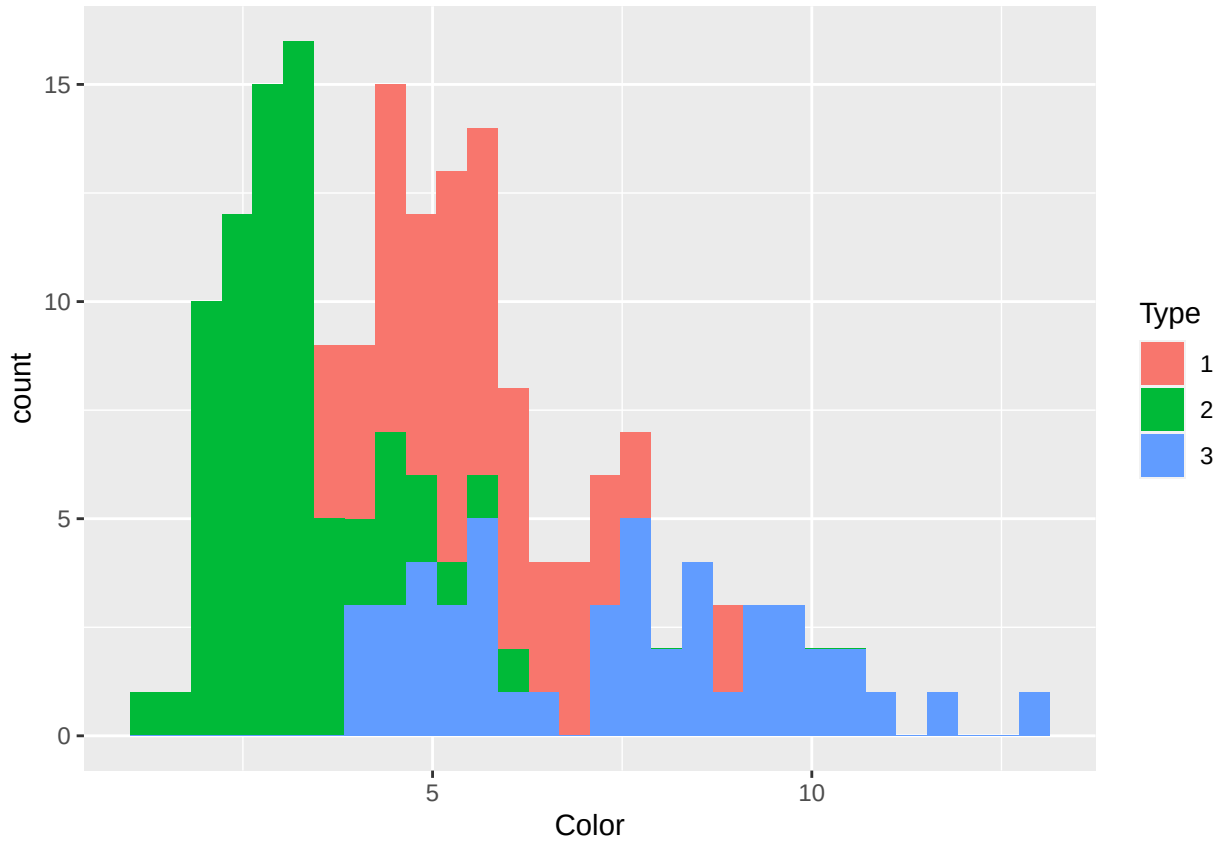
- Try out different bins for the histogram
- What is the difference of a histogram to a density plot?

```
ggplot(wine, aes(x =Color, color = Type))+
  geom_density()
```



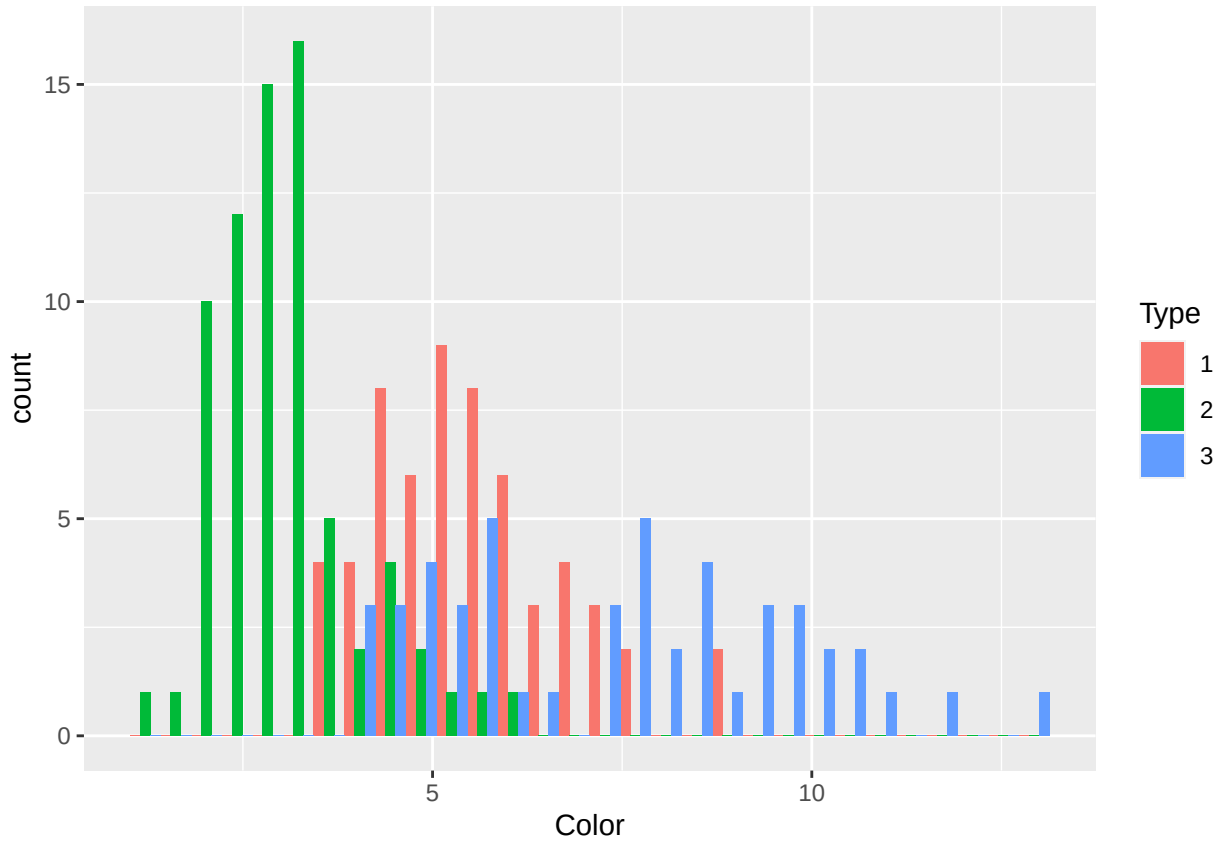

```
ggplot(wine, aes(x =Color))+
  geom_histogram(aes( fill = Type))
```

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

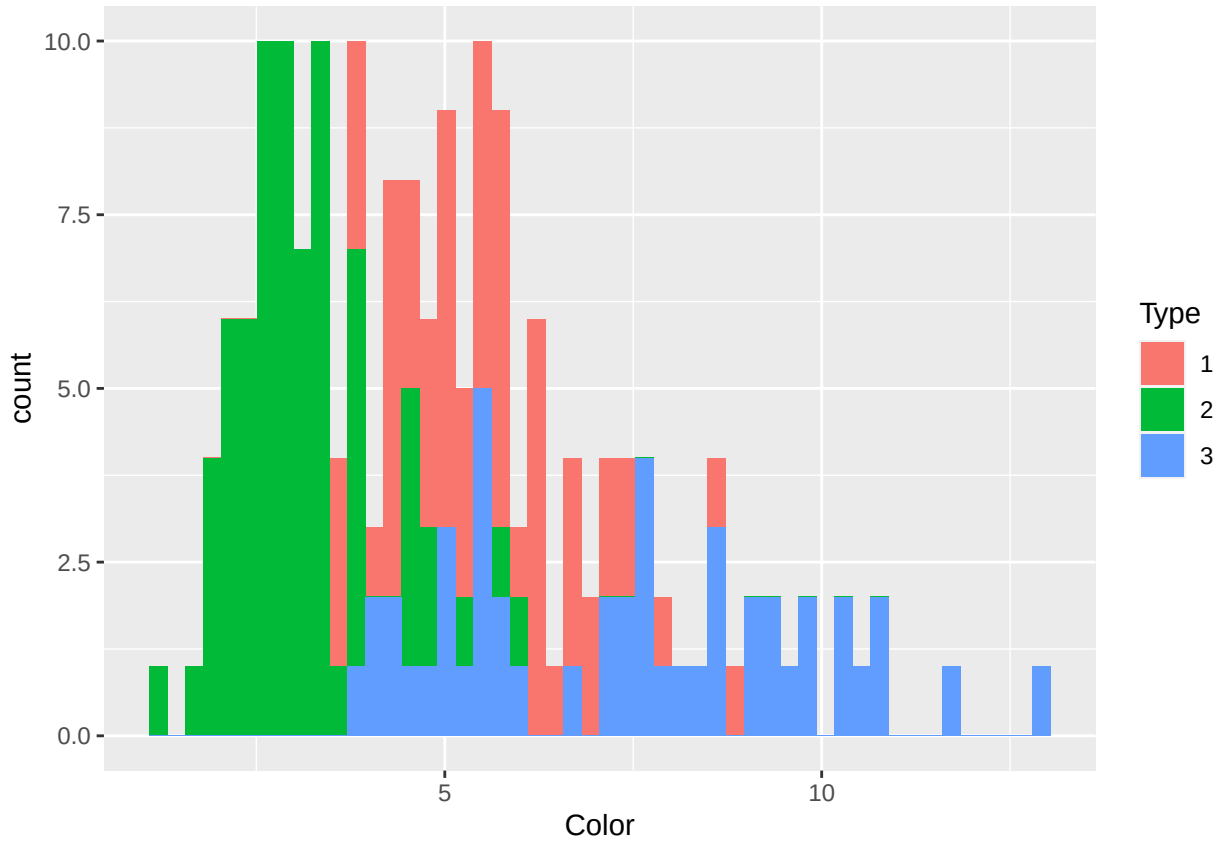



```
ggplot(wine, aes(x =Color))+
  geom_histogram(aes( fill = Type), position = "dodge")
```

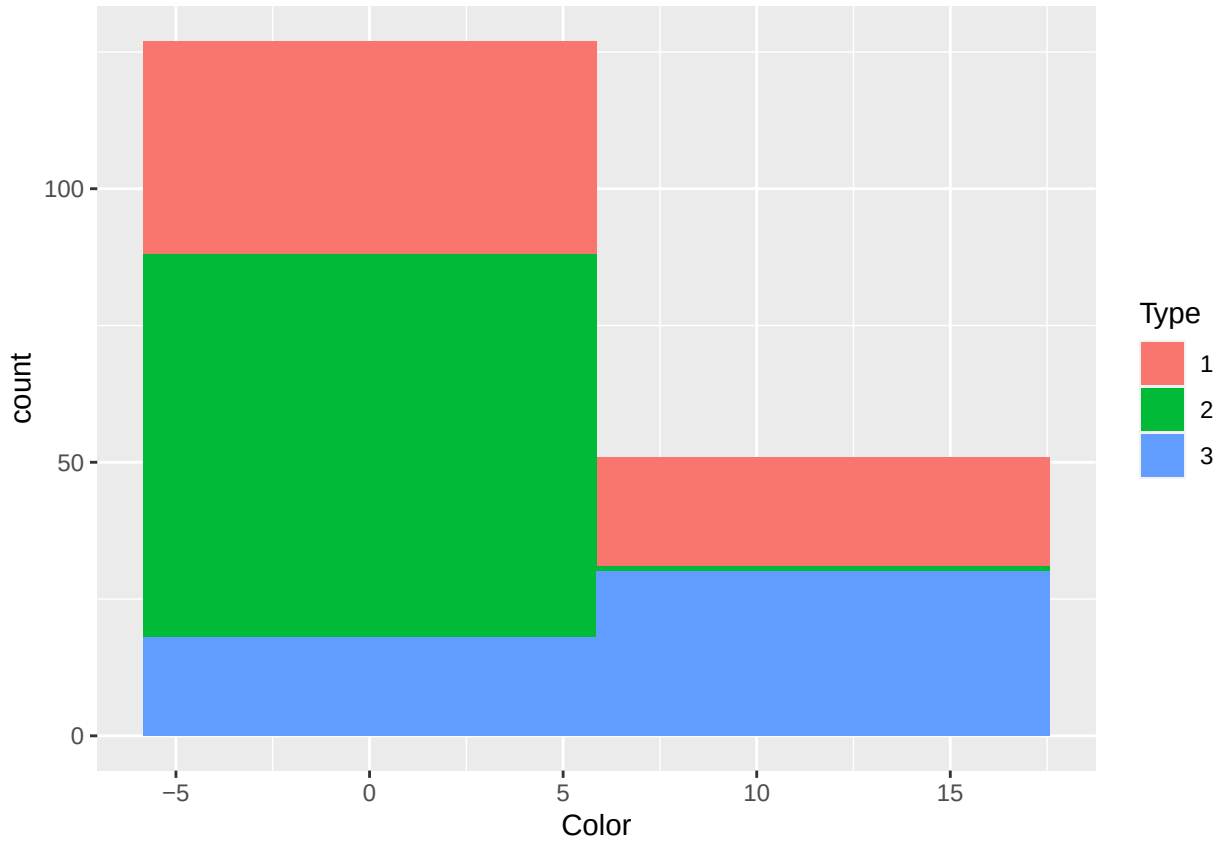
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```




```
ggplot(wine, aes(x =Color))+
  geom_histogram(aes(fill = Type), bins = 50)
```

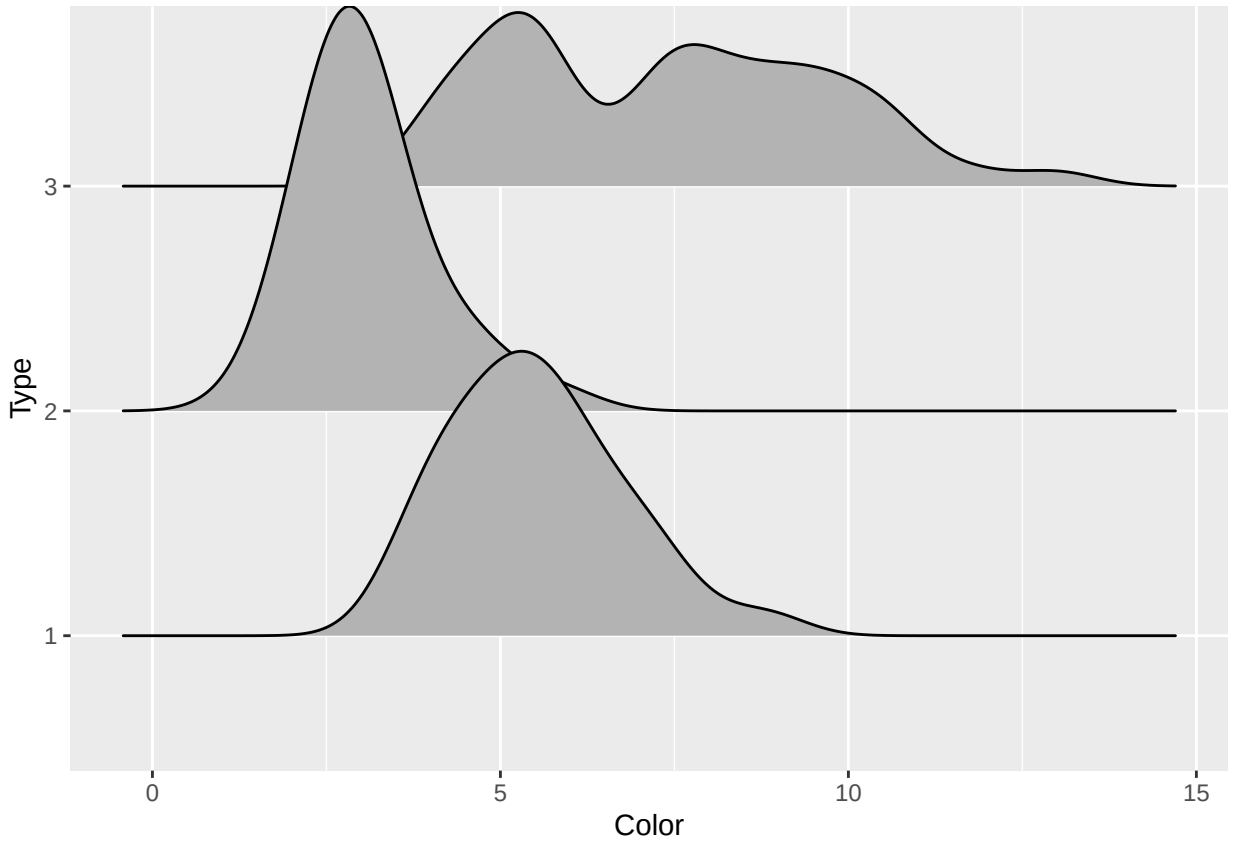



```
ggplot(wine, aes(x =Color))+
  geom_histogram(aes(fill = Type), bins = 2)
```




```
ggplot(wine, aes(x =Color, y = Type))+
  ggridges::geom_density_ridges()
```

```
## Picking joint bandwidth of 0.567
```



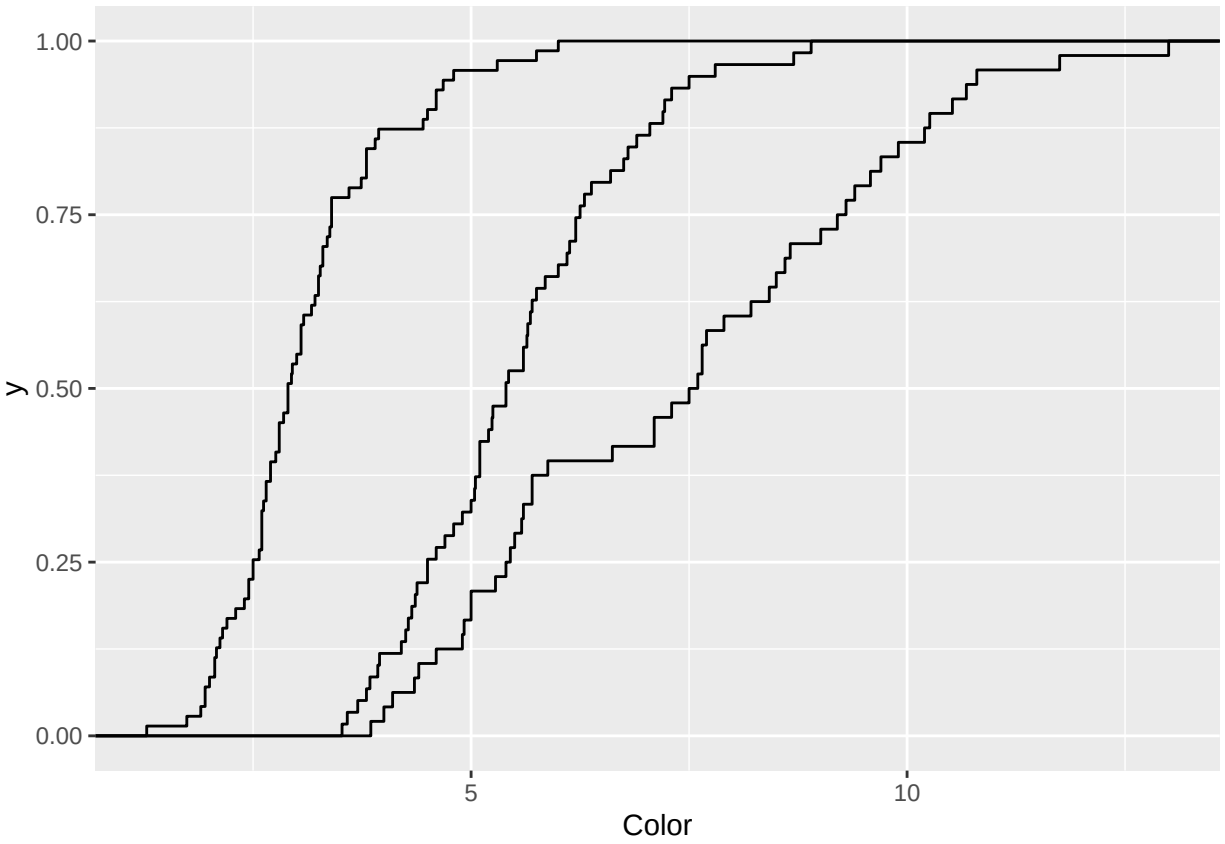
2.2.2 geom_boxplot(), geom_jitter(), geom_violin()

- How do these plots relate to a classical density plot?
- What information do you lose in which plot?
- You can also combine these three or two of them

2.2.3 stat_ecdf()

- What does this plot show you?

```
ggplot(wine, aes(x =Color, fill = Type))+
  stat_ecdf()
```



3 Statistical tests to compare distributions (of continous values)

3.1 What is a p-value? How would you explain it?

3.2 compare_means(), stat_compare_means()

- Compare_means() statisticaly compares two values
- stat_compare_means() can be added directly to a ggplot with plot + stat_compare_means() and adds the p-values to the plot

3.2.1 make a box or violin plot from above but only comparing wines of type 1 and 2 and test both functions

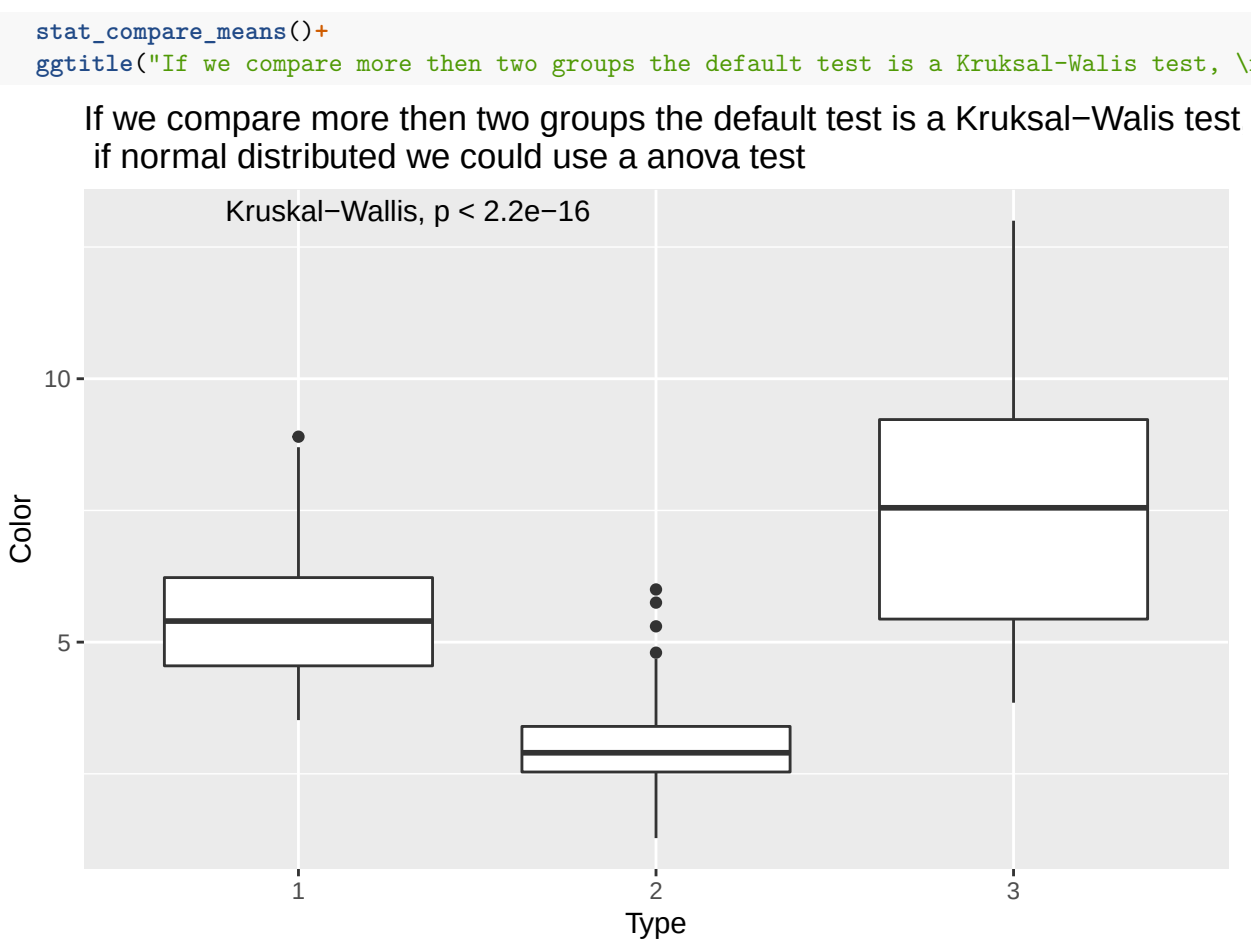
- Which statistical test is used by default?
- How do the p-values change, if you use method = "t-test"
- When would you use "t-test" as method instead?
- Use ggqqplot() and the shapiro.test() to decide if you could use a t-test here

3.3 Pairwise comparisons for three groups

- If you now make a box/violin plot with all three types of wine, what is the default test? Why is the default different now?
- Instead you can do three pairwise comparisons:
- make a list of your comparisons: my_comparisons <- list(c("3", "1"), c("1", "2"), c("3", "2"))
- use stat_compare_means(comparisons = my_comparisons)

```
ggplot(wine, aes(x =Type, y = Color))+
  geom_boxplot()+
```

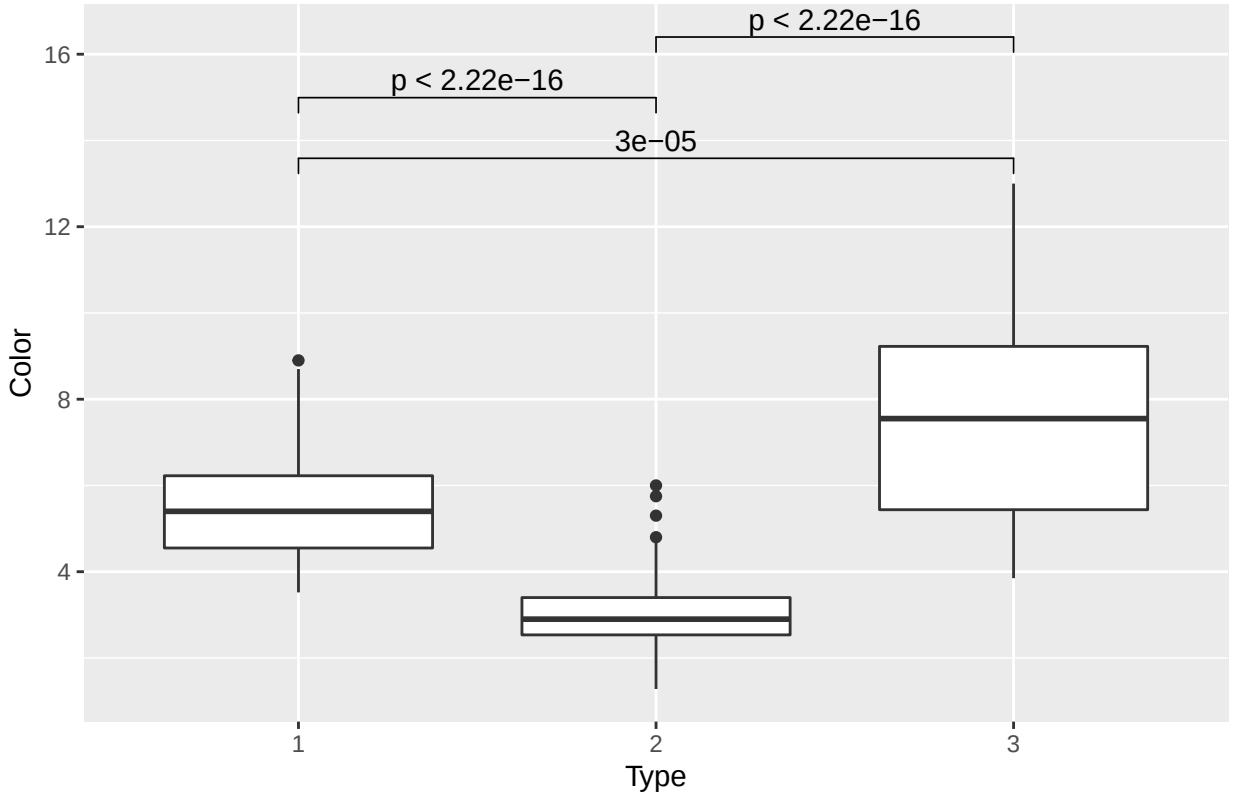
```
  stat_compare_means()+
  ggtitle("If we compare more then two groups the default test is a Kruksal-Walis test, \n if normal dis
```




```
my_comparisons <- list( c("3", "1"), c("1", "2"), c("3", "2") )

ggplot(wine, aes(x =Type, y = Color))+
  geom_boxplot()+
  stat_compare_means(comparisons = my_comparisons)+
  ggtitle("We can compare pairs of to groups instead again with Wilcoxon ")
```

We can compare pairs of to groups instead again with Wilcoxon




```
ggplot(wine, aes(x =Type, y = Color))+
  geom_boxplot()+
  stat_compare_means(comparisons = my_comparisons, method = "t.test")+
  ggtitle("Or with the t.test if we have a normal distribution \n (we saw before that thats not the cas
```

Or with the t.test if we have a normal distribution
(we saw before that thats not the case here)