

Graphical summary of categorical response

Statistics III: Multivariate Data and Regression

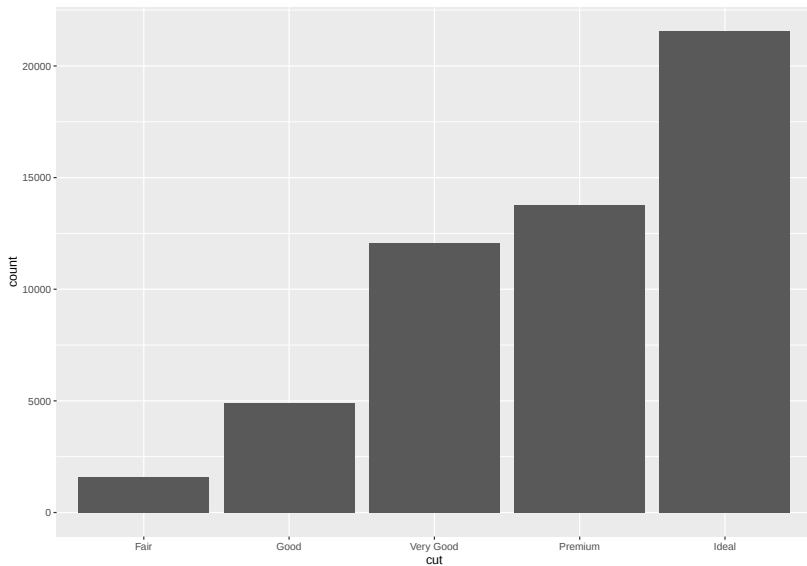
2025-11-03

Bar plot

- ▶ Bar plots visualize how a single variable is distributed by providing a count of how many times each value (discrete) of the variable is attained.

Bar plot of counts

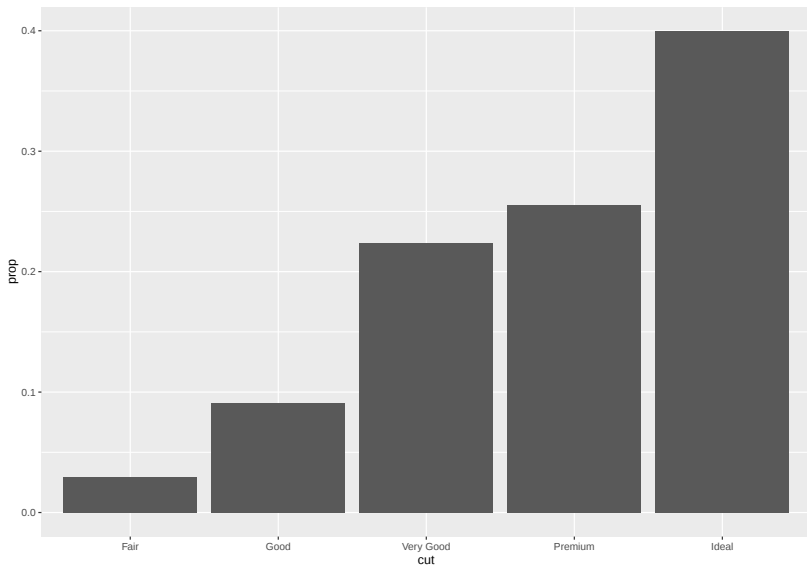
```
#install.packages("CatDataAnalysis")  
#library("CatDataAnalysis")  
library(ggplot2)  
bar1<-ggplot(data = diamonds) +  
  geom_bar(mapping = aes(x = cut))
```



Bar graph to show the proportions

```
bar2<-ggplot(data = diamonds) +  
  geom_bar(mapping = aes(x = cut, y = after_stat(prop),  
    group = 1))
```

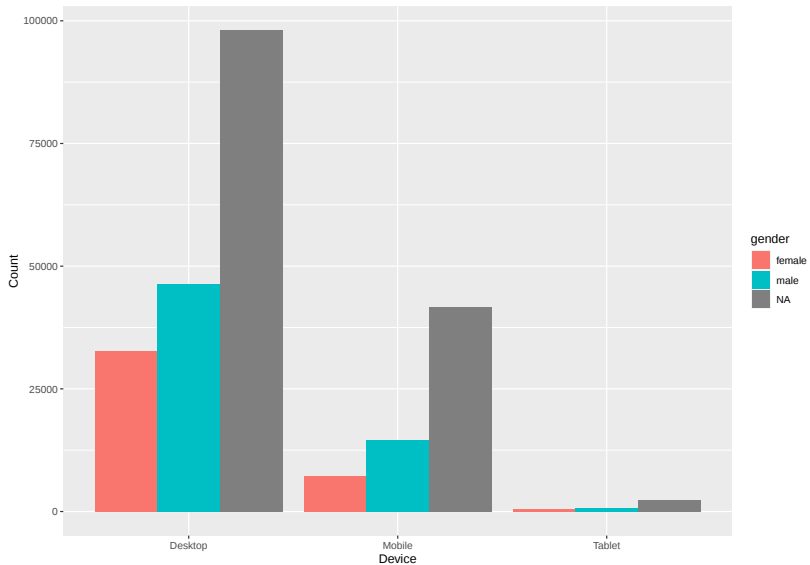
- ▶ Other alternatives are pie chart, donut chart and treemap.



Grouped bar plot

- ▶ Plot values for two levels of a categorical variable instead of one.
- ▶ You should use grouped bar chart when making comparisons across different categories of data.
- ▶ Use it when you want to look at how the second category variable changes within each level of the first and vice versa.
- ▶ So this is a situation where we have one categorical predictor and one categorical response.
- ▶ In the following we have used the type of device as predictor and gender as response. Which gender is using more laptops?
- ▶ One can look at the other way around depending on the scientific questions. Which device do men prefer?

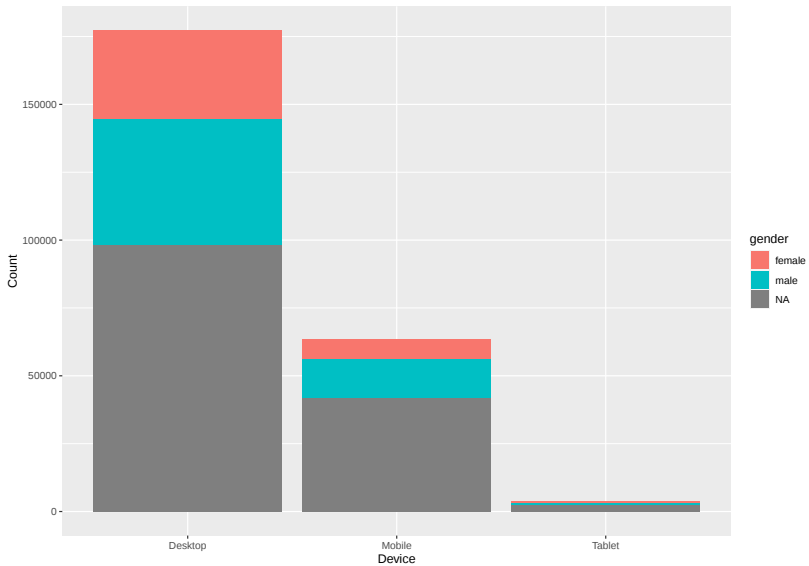
```
# https://github.com/rsquaredacademy-education/  
# online-courses/tree/master/categorical-data-in-r  
data <- readRDS('analytics.rds')  
bar3<-ggplot(data) +  
  geom_bar(aes(x = device, fill = gender),  
  position = "dodge") +  
  xlab("Device") + ylab("Count")
```



Stacked bar plots

- ▶ The bars are stacked on top of each other instead of placing them next to each other. Use stacked bar plots while looking at cumulative value.

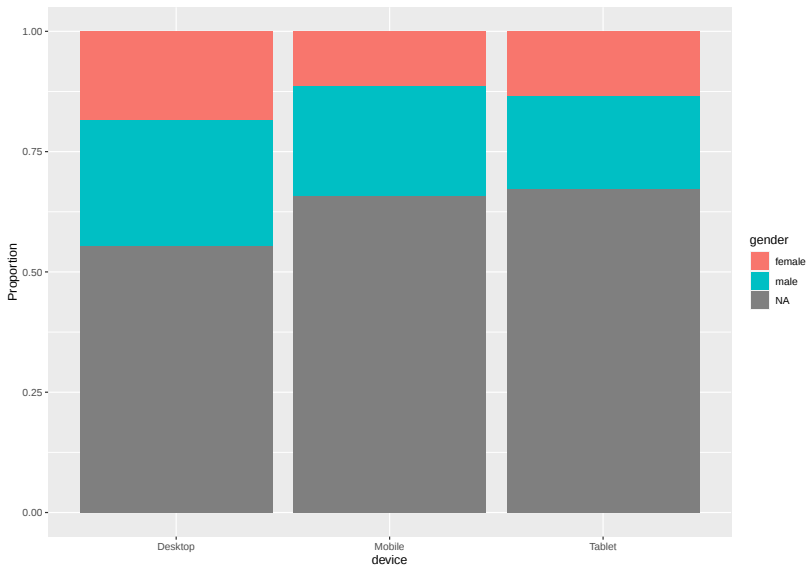
```
bar4<- ggplot(data) +  
  geom_bar(aes(x = device, fill = gender)) +  
  xlab("Device") + ylab("Count")
```



Stacked proportional bar plot

- ▶ The height of all bars in this plot are the same.
- ▶ The distribution of the second categorical variable is scaled to 1.
- ▶ The length of each bar is determined by its share in the category.
- ▶ Use this when you want to concurrently observe each of several variables as they fluctuate and as their proportions change.

```
bar5<- ggplot(data, aes(x = device, fill = gender)) +  
  geom_bar(position = "fill") +  
  labs(y = "Proportion")
```



Tile/Mosaic chart

- ▶ Useful for multiple predictors.
- ▶ In the Titanic example there are three predictors (gender, child/adult and class) and one response (saved yes/no). All categorical.
- ▶ In the HairEyeColor example, the role of predictor/response is unclear. This is more of a multivariate data setting (not regression). There are 3 categorical variables: gender, Hair Color and Eye Color.

```
if(!require('vcd')) {  
  install.packages('vcd')  
  library('vcd')  
}
```

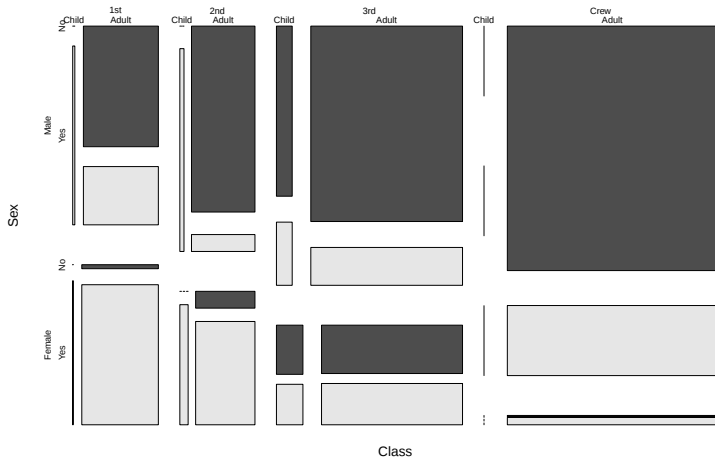
```
## Loading required package: vcd
```

```
## Loading required package: grid
```

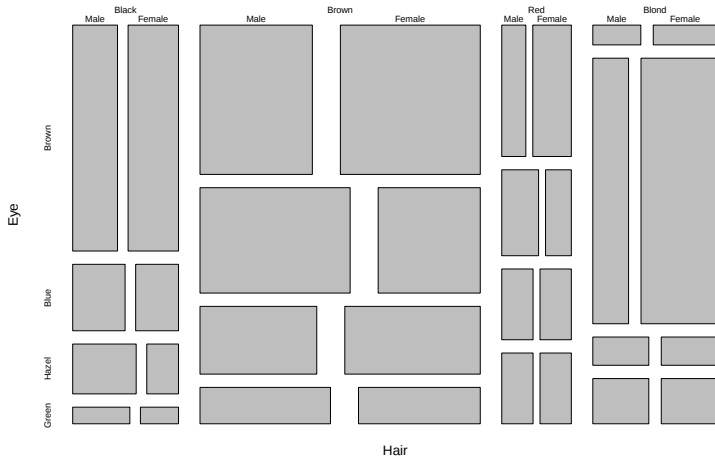
```
data("Titanic")  
mosaicplot(Titanic,color=TRUE)
```

```
mosaicplot(HairEyeColor)
```

Titanic



HairEyeColor



Parallel coordinates plot.

- ▶ Multiple quantitative responses. Categorical predictor.
- ▶ Fisher's Iris data. Predictor is species. Response are Sepal/Petal Length/Width.

```
library(MASS)
ir <- rbind(iris3[, , 1], iris3[, , 2], iris3[, , 3])
parc<-parcoord(ir, col = 1 + (0:149)%/%50)
```

