Data Science and R – Lab 10

Plotting with ggplot

`0. Loading ggplot2`

Install and load the ggplot2 library

```
install.packages.('ggplot2')
library(ggplot2)
```

Ensure the package 'ISLR' is installed and loaded.

`1. Carseats dataset`

a) Plot a scatter plot of the `Price (x)` versus `Sales (y)` variables in the `Carseats` dataset. Use `ggplot()` with `aes()` to specify the x and y parameters and `geom_point()`

____

b) Transform the y-axis values to their square root. Use `trans='sqrt'` in `scale_y_continuous()` and add it as a layer to your plot

____

c) Change the y axis label to "Sales in thousands" using the `ylab()` function added to your plot layer

____

d) Add a regression line using the `geom_smooth()` function with `method=lm`. Within `geom_smooth()`, set the line's color to red using col and do not show the confidence interval by setting `se=F`

____

e) Split the data by colors to denote different `Urban`. Is there a pattern between `Urban` and `Sales` or `Urban` and `Price`?

____

f) Split the data by color again but this time using the `US` variable. You could also vary the shape of distinct US values by adding `shape=as.factor(US)` in addition to color in `aes()`. Is there a relationship between the presence of US with `Price` or `Sales`?

____

g) Split the data by color again but this time using the `ShelveLoc` variable. You could also vary the shape of distinct quality of the shelving location values by adding `shape=ShelveLoc` in addition to color in `aes()`. Is there a relationship between the shelving locations with `Price` or `Sales`?

____

2. Iris dataset

a) Plot the `iris` dataset's `Petal.Width` against `Sepal.Length`. Can you see two distinct areas in the plot? We are going to try to understand that.

____

b) Add a third variable `Species` to this plot. It should cut the data by `color`. Can you explain what this says about the sepal width and species type or petal length with species type?

____

c) Cut the data again using `size` to denote different `Petal.Length`. Keep the previous cut using species in different colours. Now you can see 4 different variables in one graph. What can you explain about the relationship between these 4 variables?

____

3. College dataset

a) Plot the number of applicants, `Apps` against the the number of new students enrolled, `Enroll`.

____

b) The plot may not be easy to see because many dots are clumped at the bottom left part. We will ease visualisation by transforming both the `x` and `y` axes to their logs. Replot with these transformations. Are there generally more students enrolled if there are more applicants?

____

c) We want to see if there is a difference between private and public colleges. Cut the plot above by the type of college, either private or public (`Private`). Which type of college has more applications in general?

____

d) We now want to investigate the costs. Plot the tution fees (`Outstate`) against room and boarding costs (`Room.Board`). This time, cut the plot with the cost of `Books`. Is the cost of books higher for college with higher fees?

____

e) Finally we want to see if different types of colleges have different costs. Cut the plot in c) by `Private`. Are private universities more expensive in terms of tuition and boarding?

____