Lab Session 7

Graphical Analysis with ggplot2()

Introduction

ggplot2 is a powerful R package, implemented by Hadley Wickham, for producing nice graphs. The ggplot2 package allows you to build very complex graphs layer by layer. Unlike graphs we construct using the base functions in R, ggplot2 takes care of details like legends and choice of plotting symbols automatically, although you can customise these choices if you wish.

Install.packages("ggplot2")

library(ggplot2)

Basic Plotting with ggplot

The ggplot design is very elegant, takes some thinking to get used to, but is extremely powerful. The central premise is to characterize the building pieces behind ggplot plots as follows:

- Data, a data frame of entities and attributes.
- The mapping between data attributes and graphical (aesthetic) characteristics

Some of the graphical characteristics we will commonly map attributes to include:

Argument	Definition
`x`	position along x axis
`y`	position along y axis
`color`	color
`shape`	shape (applicable to e.g., points)
`size`	size
`label`	string used as label (applicable to text)

- The geometric representation of these graphical characteristics.

  We can include multiple geometric representations in a single plot, for example points and text, by adding (+) multiple geom_<representation> functions. Also, we can include mappings inside a geom_ call to map characteristics to attributes strictly for that specific representation.

Representations we will use frequently are

Function	Representation
geom_point	points
geom_bar	rectangles
geom_text	strings
geom_smooth	smoothed line (advanced)
geom_hex	hexagonal binning

```
# NOTE: this is pseudo-code. It will not run!

ggplot(data = <DATA>) +
  <GEOM_FUNCTION>(mapping = aes(<MAPPINGS>))
```

Let's see how we can apply these for an R built-in data frame named mtcars. The data set belongs to the dplyr package and has to be pre-loaded into the R workspace prior to its use.

Bar Graph

```
library(dplyr)
data("mtcars")


#label the catergorical varibles
mtcars2 <- within(mtcars, {
  vs <- factor(vs, labels = c("V", "S"))
  am <- factor(am, labels = c("automatic", "manual"))
  cyl  <- ordered(cyl)
  gear <- ordered(gear)
  carb <- ordered(carb)
})

#Bar chart
ggplot(data=mtcars, mapping=aes(x=cyl))+geom_bar(color='blue', fill='blue')
```

Horizontal bar plot

```
P<- ggplot(data=mtcars2, mapping=aes(x=cyl))+geom_bar(color='blue', fill='blue')
P+coord_flip()
```

Ordered Bar plot

```
ggplot(data=mtcars2, mapping=aes(x=reorder(vs,vs,function(x)
length(x))))+geom_bar(stat="count",color='blue', fill='blue')
```

ggplot2 is probably the best option to build multiple and stacked barchart. The input data frame requires to have two categorical variables that will be passed to the x and fill arguments of the aes() function.

Multiple Bar Charts

```
ggplot(mtcars2,aes(x=am, fill=vs))+geom_bar(position="dodge")
```

Stacked Bar Charts

```
ggplot(mtcars2,aes(x=am, fill=vs))+geom_bar(position="stack")
```

Perncentage Component Bar Charts

```
ggplot(mtcars2,aes(x=am, fill=vs))+geom_bar(position="fill")
```

Customize graphs

As usual, some customization are often necessary to make the chart look better and personnal. Let's:

- add a title
- use a theme (theme_ *: there are lot of available themes to select)
- change color palette. See more here.
- customize axis titles

Add Title and themes

```
ggplot(mtcars2,aes(x=am, fill=vs))+geom_bar(position="fill")

+ labs(x = "Transmission", y = "Percentage", title = "Relationship between Car
Engine type and transmission", colour="Engine")

+theme_gray() #adding a theme for graphs
```

Quantitative Data

.

Dot plot

```
ggplot(data =mtcars2 , mapping = aes(x=gear))+geom_dotplot()
```

#Use a categorical variable to colour dots

```
ggplot(data =mtcars2 , mapping = aes(x=gear,fill=vs))+geom_dotplot()
```

Histogram

```
ggplot(data =mtcars2 , mapping =
aes(x=mpg))+geom_histogram(fill="blue",color="black")
```

#Use bin size to get a better histogram

```
ggplot(data =mtcars2 , mapping =
aes(x=mpg))+geom_histogram(fill="blue",color="black",bins = 8)
```

#Use a categorical variable to see variations on the histogram

```
ggplot(data =mtcars2, mapping = aes(x=mpg,fill=vs))+geom_histogram(bins = 8)
```

You probably want to make the bars semi-transparent using position = "identity"so you can can distinguish the overlapping data.

```
ggplot(data =mtcars2, mapping = aes(x=mpg,fill=vs))+geom_histogram(bins =7,
position = "identity", alpha = 0.4)
```

Yet another way to represent the histograms for different categories is to use faceting,

```
ggplot(data =mtcars2, mapping = aes(x=mpg,fill=vs))+geom_histogram(bins
=7)+facet_wrap(~vs, ncol = 1)
```

Denisty plot

```
ggplot(data =mtcars2 , mapping = aes(x=mpg))+geom_density()
```

#Use a categorical variable to get separate density plots

```
ggplot(data =mtcars2 , mapping = aes(x=mpg, fill=vs))+geom_density(alpha=0.25
)
```

Box plot

Boxplots provide a compact summary of single variables, and are most often used for comparing distributions between groups.

```
ggplot(data =mtcars2 , mapping = aes(y=mpg))+geom_boxplot()
```

Boxplots are most commonly drawn with the categorical variable on the x-axis.

```
ggplot(data =mtcars2 , mapping = aes(x=vs, y=mpg, fill=vs))+geom_boxplot()
```

Scatter plot

```
ggplot(data=mtcars2,aes(x=wt,y=mpg))+geom_point()
```

Scatter plot with smooth curves

```
ggplot(data=mtcars2,aes(x=wt,y=mpg))+geom_point() + geom_smooth()
```

Scatter plot with Categorical Information

```
ggplot(data=mtcars2,aes(x=wt,y=mpg))+geom_point(aes(color = vs, shape = vs), size = 2, alpha = 0.6)
```

Density plot - Two variables

```
ggplot(data=mtcars2,aes(x=wt,y=mpg))+geom_density2d()
```

Density plot with Categorical Information

```
ggplot(data=mtcars2,aes(x=wt,y=mpg))+geom_density_2d(aes(color = vs))
```

When you draw plots using ggplot2 pacage always use labs to include title, axis titles etc.

Lab Exercise

In this lab, you will determine which factors are associated with low birth weight, among other explorations.

- The dataset is named birthweight.xls

Data:

Column	Abbreviation	Variable
C1	ID	Identification Code
C2	LOW	Low Birth Weight (no = Birth Weight >= 2500g,yes = Birth Weight < 2500g)
C3	AGE	Age of the Mother in Years
C4	LWT	Weight in Pounds at the Last Menstrual Period
C5	RACE	Race (1 = White, 2 = Black, 3 = Other)
C6	SMOKE	Smoking Status During Pregnancy (Yes, No)
C7	PTL	History of Premature Labor (0 = None 1 = One, etc.)
C8	HT	History of Hypertension (Yes, No)
C9	UI	Presence of Uterine Irritability (Yes, No)
C10	FTV	Number of Physician Visits During the First Trimester
C11	BWT	Birth Weight in Grams

Use ggplot2 and perform univariate and bivariate analysis for the above dataset.