Graphs using ggplot()

James Mburu

February 17, 2016

Graphical exploration of data.

- To have an understanding of you data we normally conduct exploratory data analysis (EDA) which can be graphical or numerical.

- Primarily EDA is for seeing what the data can tell us before the formal modeling or hypothesis testing task.

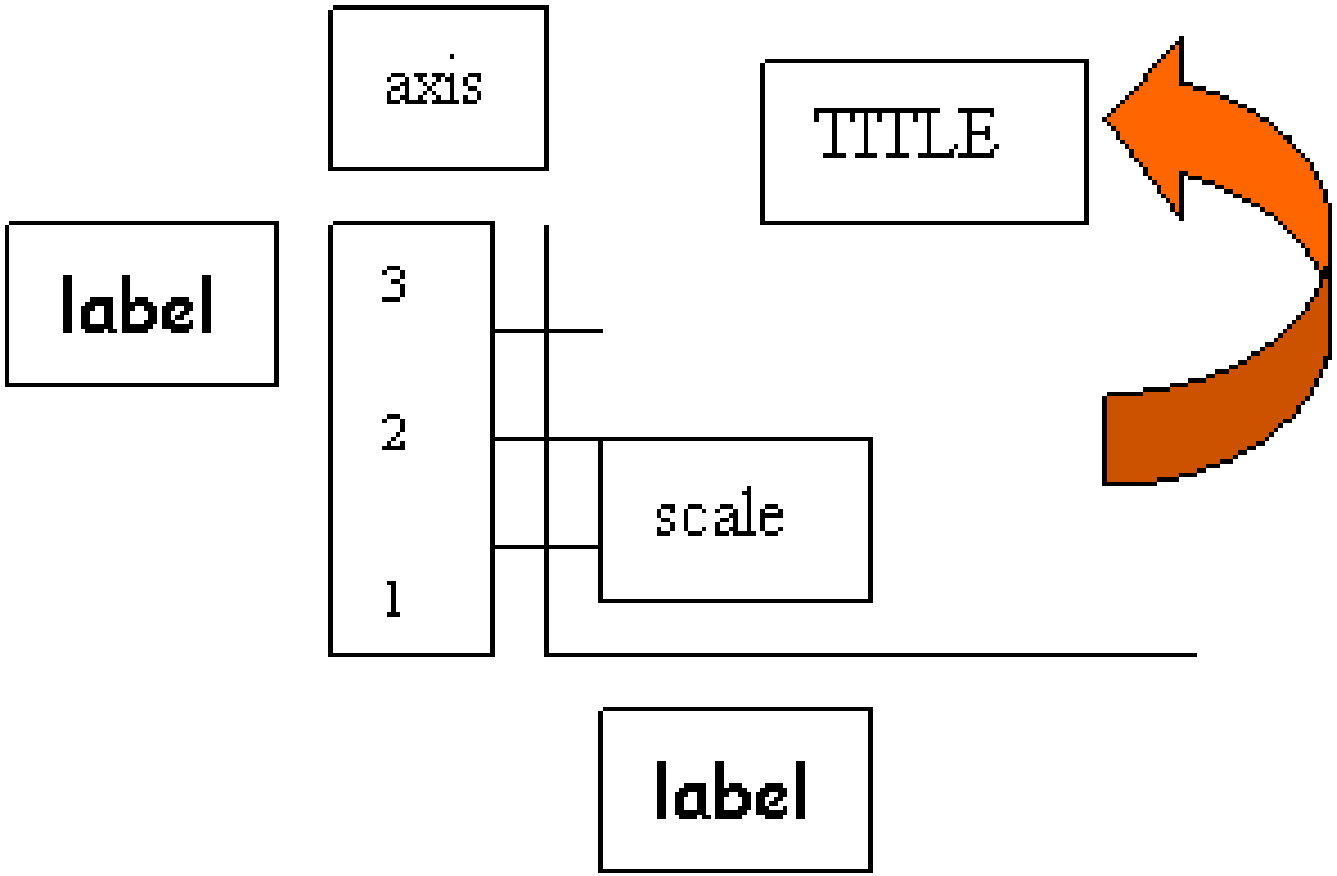
- Typical graphical techniques used in EDA are:
  - Scatter plots,
  - Box plots,
  - Bar plots.

Scatter plot

Scatter plot

- Its a useful summary of a set of bivariate data (two variables)
- It pairs up values of two quantitative variables in a data set with the aim of giving a good visual picture of the relationship between the two variables.
- The resulting pattern indicates the type and strength of the relationship between the two variables.
- Usually drawn before working out a linear correlation coefficient or fitting a regression line.

Parts of a graph



Introduction to ggplot()

- **ggplot** - R package for data exploration and producing plots.
- It produces fantastic-looking statistical graphics.
  author- Hadley Wickham
- Get the package:

```
install.packages("ggplot2") # To install the package
library(ggplot2) # To load the package
```

Introduction to ggplot() . . .

- ggplot2 provides two ways to produce plot objects:
  - qplot() - quick plot
    - designed to be very similar to plot() and simple to use
    - may make it easy to produce basic graphs
  - ggplot() - grammar of graphics plot
    - a bit challenging BUT allows much more flexibility when building graphs

Components of the Graphics in ggplot2.

- **Data**: must be stored as an R data frame

- **Coordinate system**: describes 2-D space that data is projected onto
  - e.g. Cartesian coordinates, map projections, ...

- **Geoms**: short for geometric objects, describe the type of plot you will produce.
  - e.g. points, lines, bar, ...
  - geom_point, geom_line, geom_bar, geom_boxplot, ...

Components of the Graphics in ggplot2 . . .

- **Aesthetics**: describe visual characteristics that represent data
  - e.g. size, color, shape, fill, line type, ...

- **Stats**: describe statistical transformations that typically summarize data
  - e.g. counts, means, medians, regression lines, ...

- **Facets**: help display subsets of the dataset in different panels.

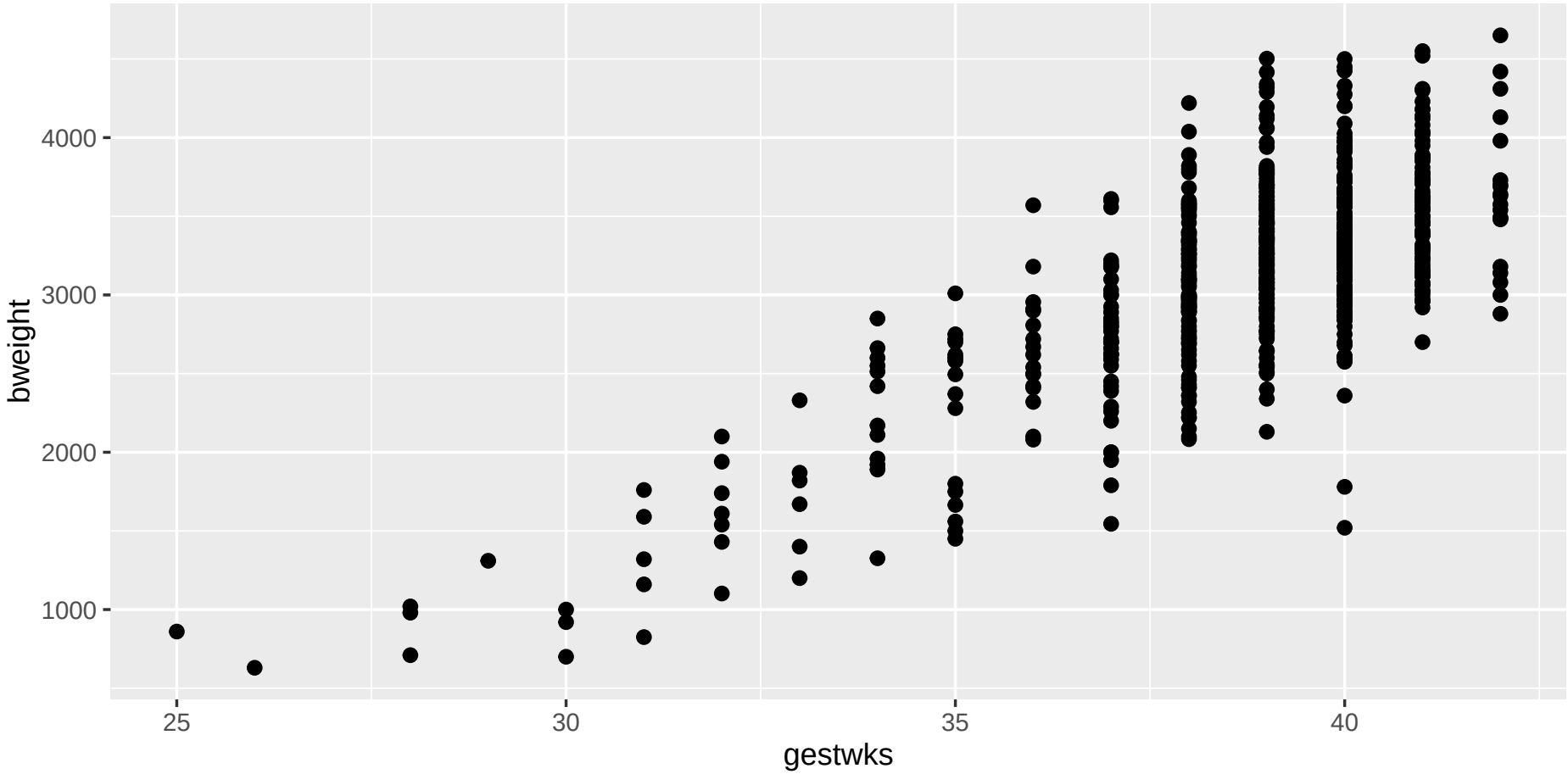
- **Annotation**: Specialised functions for adding annotations to a plot.

Data

id	matage	ht	gestwks	sex	bweight	lbw	agegrp
1	33	2	38	Female	2410	Weight<2500	30-34 yrs
2	34	2	39	Female	2977	Normal 2500+	30-34 yrs
3	34	2	36	Female	2100	Weight<2500	30-34 yrs
4	30	2	39	Male	3270	Normal 2500+	30-34 yrs
5	35	2	38	Female	2620	Normal 2500+	35-39 yrs
6	37	2	38	Male	3260	Normal 2500+	35-39 yrs

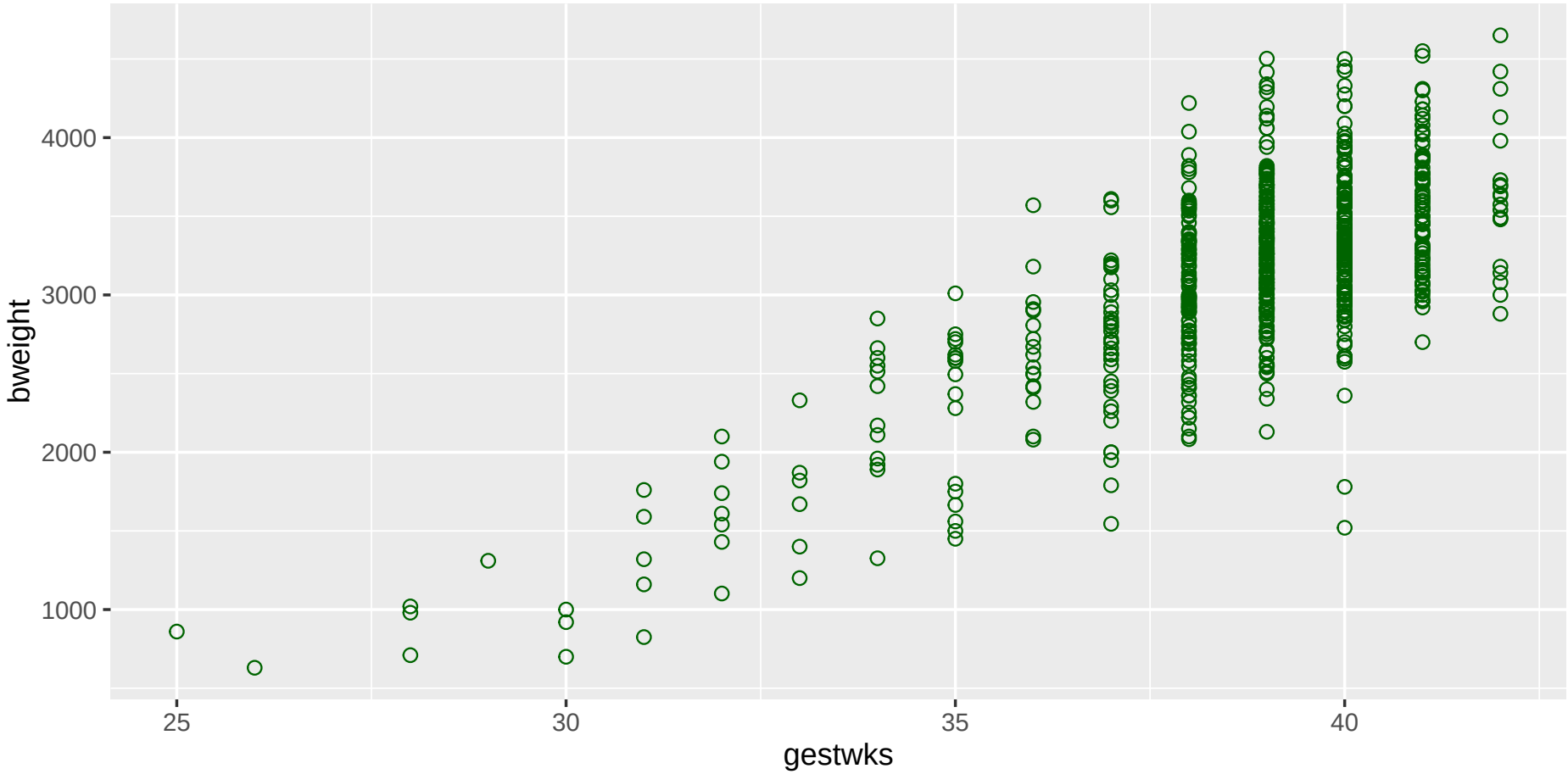
Scatter plot.

```
p <- ggplot(data=birth) #initializes a ggplot object
p + geom_point(aes(x=gestwks,y=bweight),size=2)
```



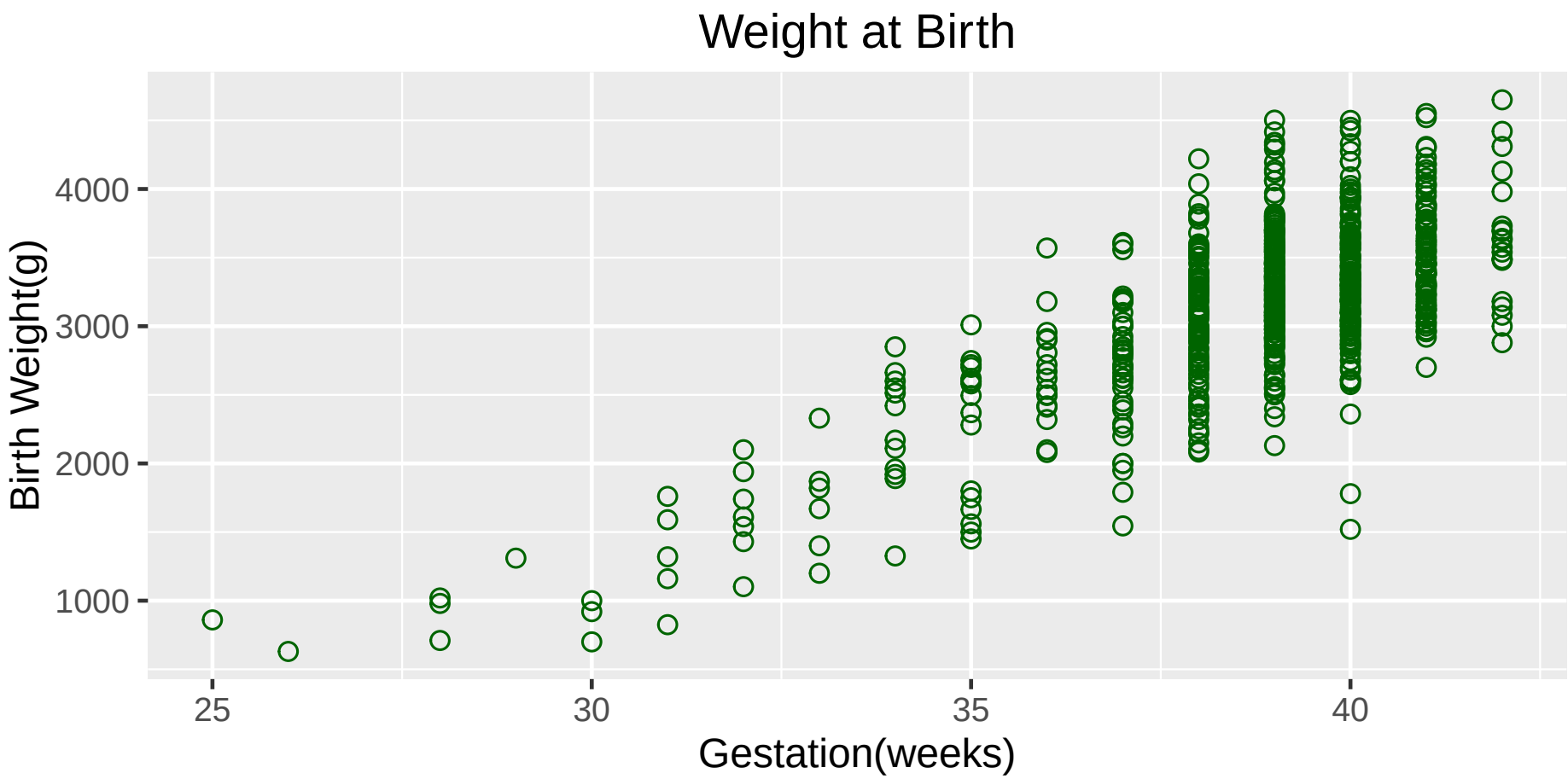
Scatter plot . . .

```
(p1<-p + geom_point(aes(x=gestwks,y=bweight),size=2,
                    color="darkgreen",shape=1))
```



Scatter plot . . .

```
(p1<- p1 + labs(title="Weight at Birth",x="Gestation(weeks)",
        y="Birth Weight(g)"))
```



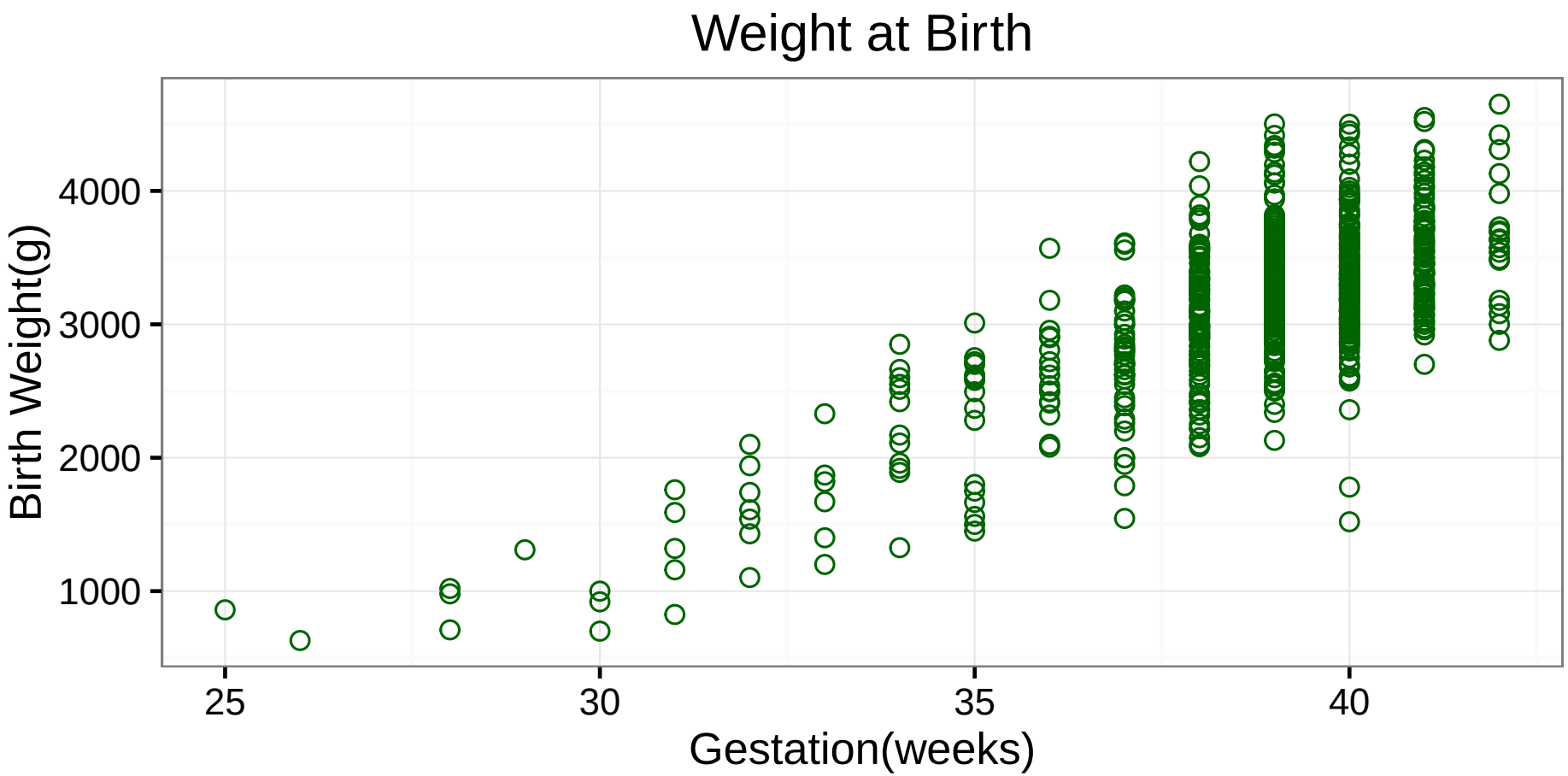
Scatter plot . . . Theme

Controls appearance.

- helps make plot visually pleasing by allowing addition/modification/deletion of;-
- titles, axis labels, tick marks, axis tick labels and legends

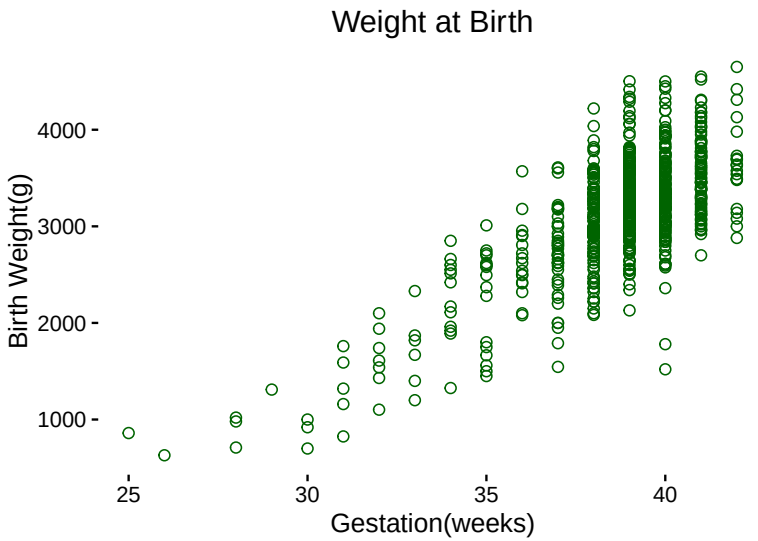
Scatter plot. . . Theme: Overall Look

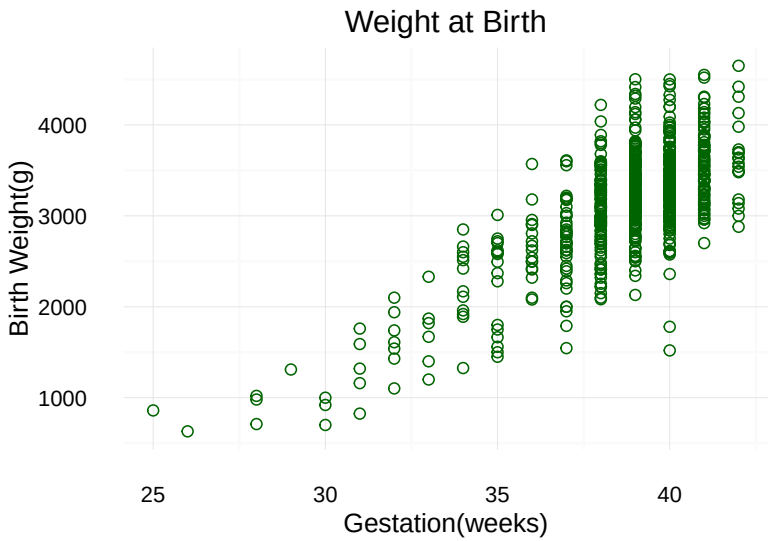
```
(p1<-p1 + theme_bw())
```

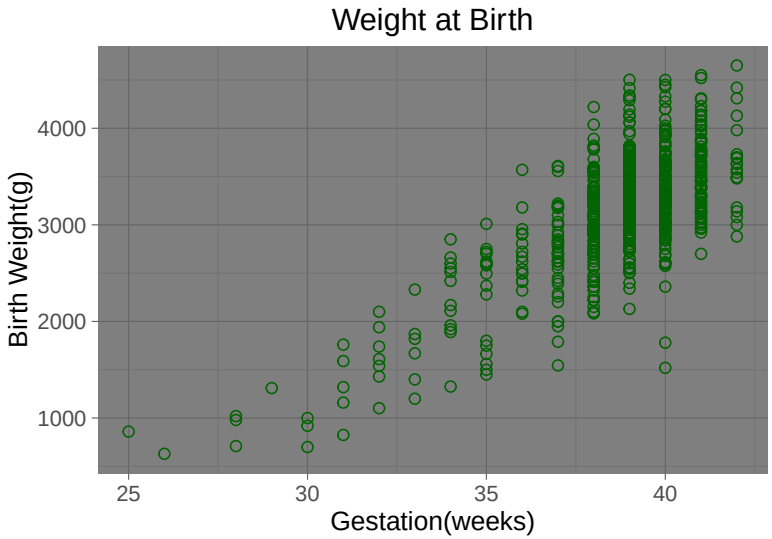


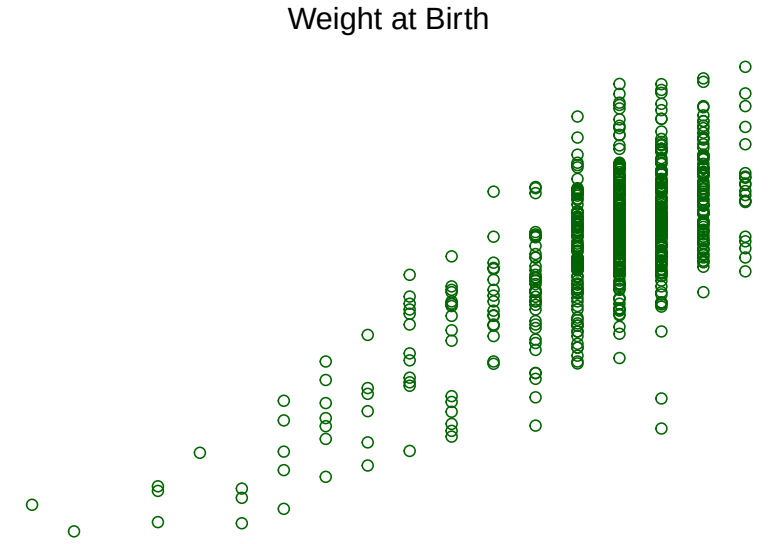
Scatter plot. . . Theme: Overall Look

```
p.1<-p1+theme_classic(); p.2<- p1+theme_minimal(); p.3<-p1+theme_dark(); p.4<-p1+theme_void()
```





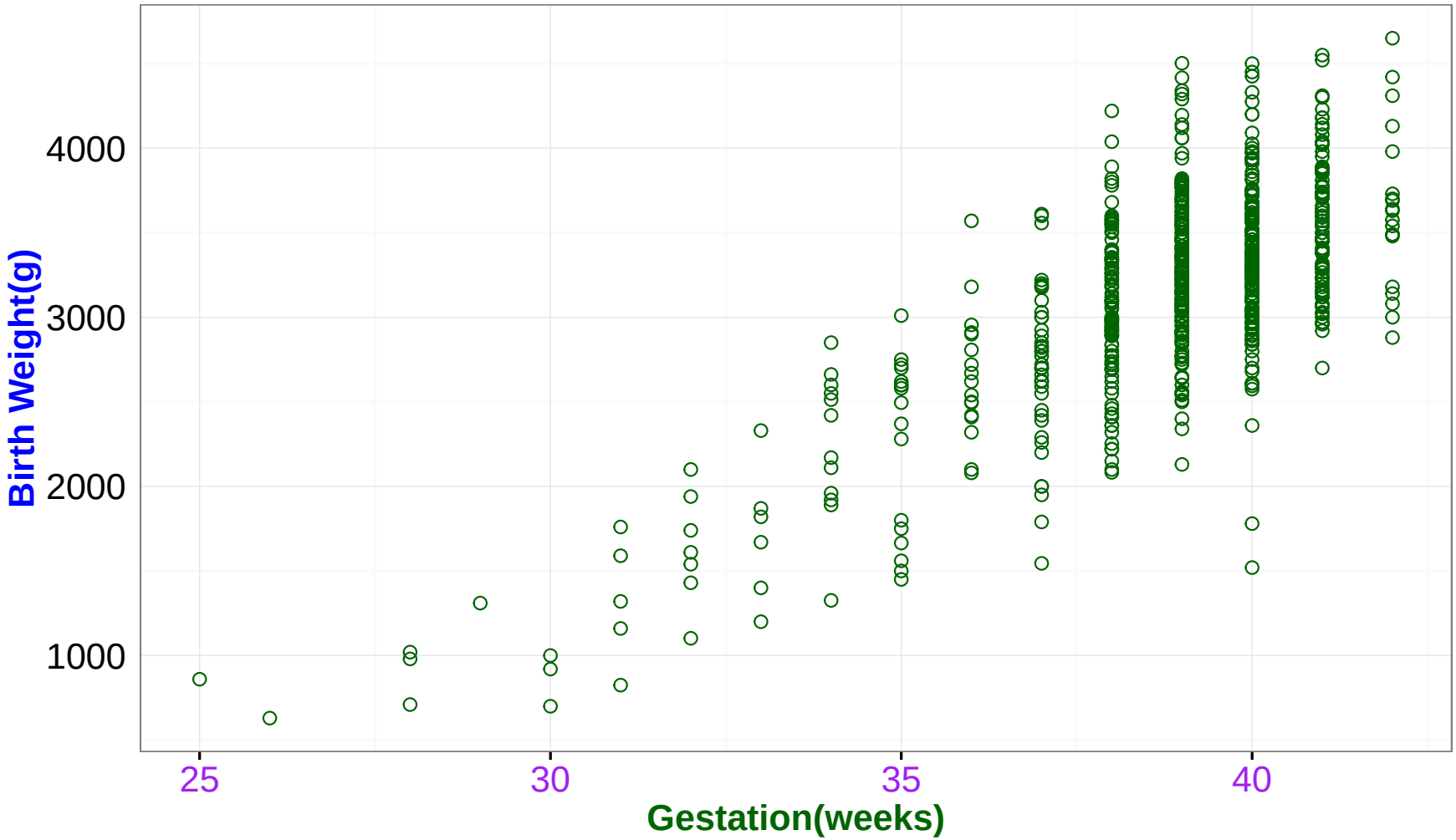




Scatter plot . . . Theme: titles, tick marks, tick labels

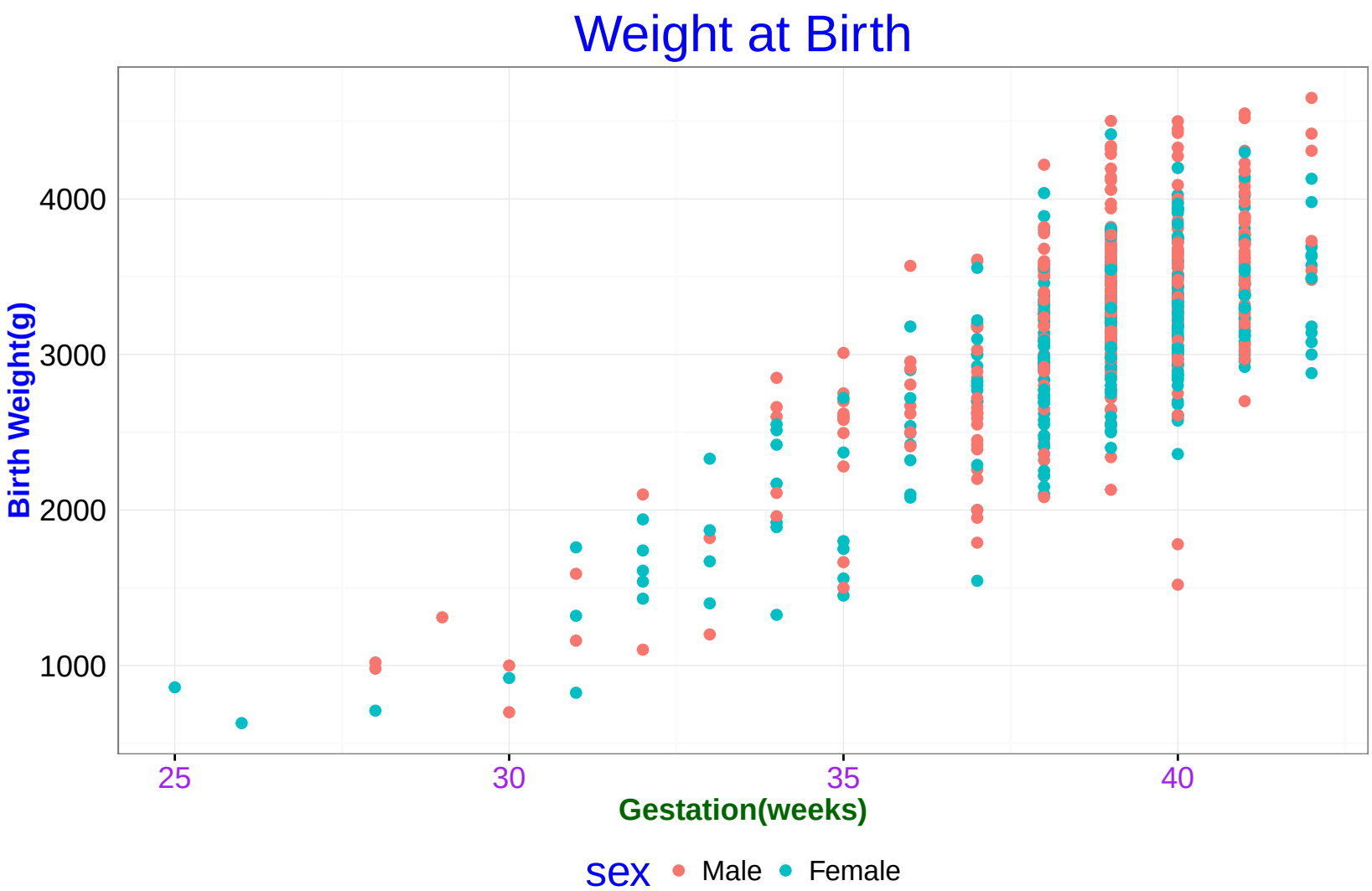
```
p1 <- p1 + theme(title=element_text(color="blue", size=20),
           axis.title=element_text(size=14,face="bold"),
           axis.title.x=element_text(color="darkgreen"),
           axis.text=element_text(size=14),
           axis.text.y=element_text(color="black"),
           axis.text.x=element_text(color="purple"),
           axis.ticks.y=element_blank())
```

Weight at Birth



Theme: Legends

```
p <- ggplot(data=birth) #initializes a ggplot object
p1<- p + geom_point(aes(x=gestwks,y=bweight,color=sex),size=2)
p1<- p1 + labs(title="Weight at Birth",x="Gestation(weeks)",
         y="Birth Weight(g)") + theme_bw()
p1<- p1 + theme(title=element_text(color="blue", size=20),
            axis.title=element_text(size=14,face="bold"),
            axis.title.x=element_text(color="darkgreen"),
            axis.text=element_text(size=14),
            axis.text.y=element_text(color="black"),
            axis.text.x=element_text(color="purple"),
            axis.ticks.y=element_blank())
p1 <- p1 + theme(legend.key=element_blank(),
            legend.text=element_text(size = rel(1.1)),
            legend.direction="horizontal",legend.position="bottom")
```

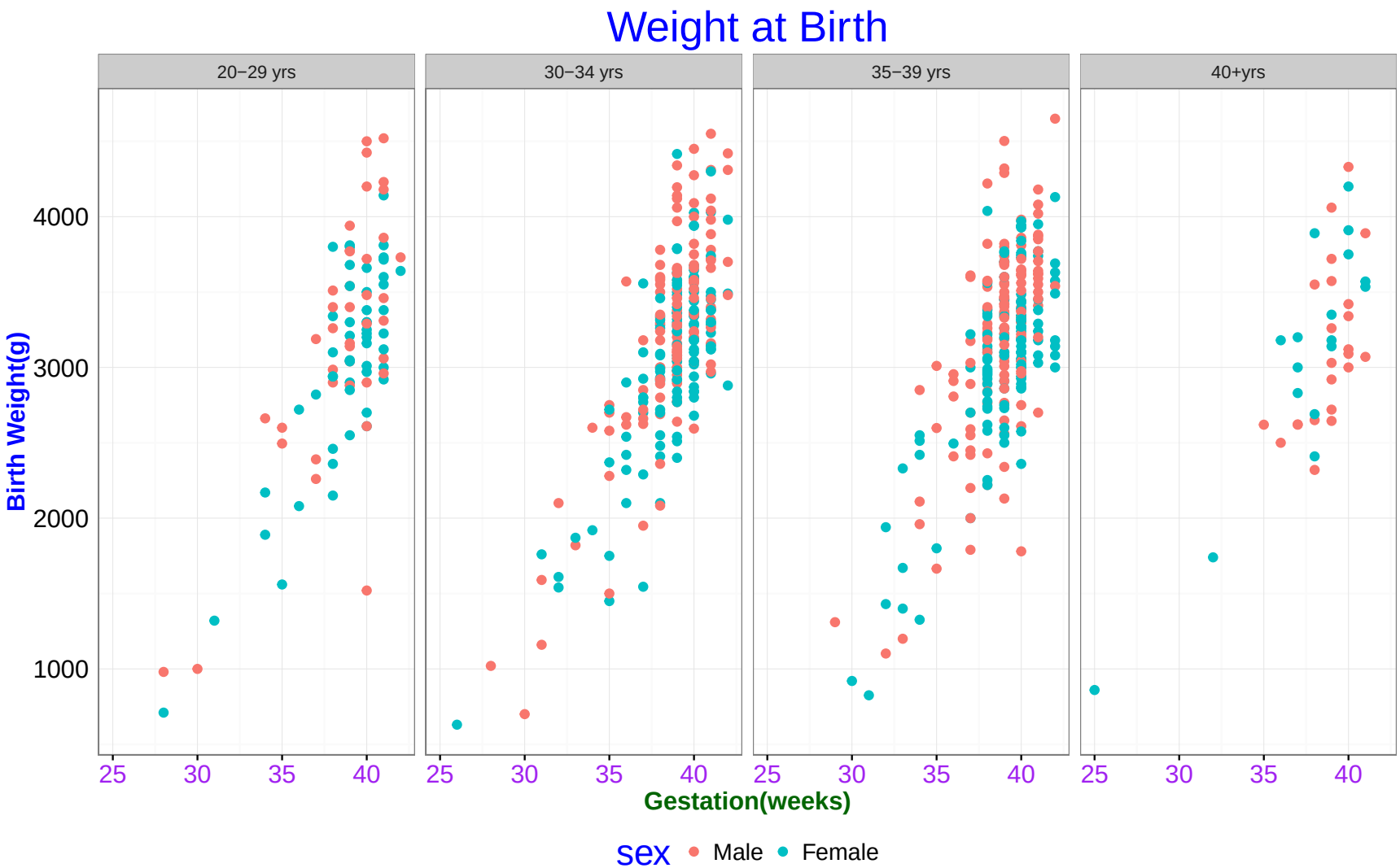

Weight at Birth
Birth Weight(g)
4000
3000
2000
1000
25
30
35
40
Gestation(weeks)
sex
Male
Female

Facets.

-. Facets display subsets of the dataset in different panels.

```
p1<- p1 + facet_grid(. ~ agegrp) #Lay out panels in a grid.
```

Facets


Weight at Birth
20−29 yrs
30−34 yrs
35−39 yrs
40+yrs
Birth Weight(g)
4000
3000
2000
1000
25
30
35
40
Gestation(weeks)
sex
Male
Female

Saving Graphs.

```
ggsave() # saves last plot displayed..Default
```

-. Formats file name extension
.pdf .jpg .png .bmp .svg .wmf .tex .tiff .ps .eps

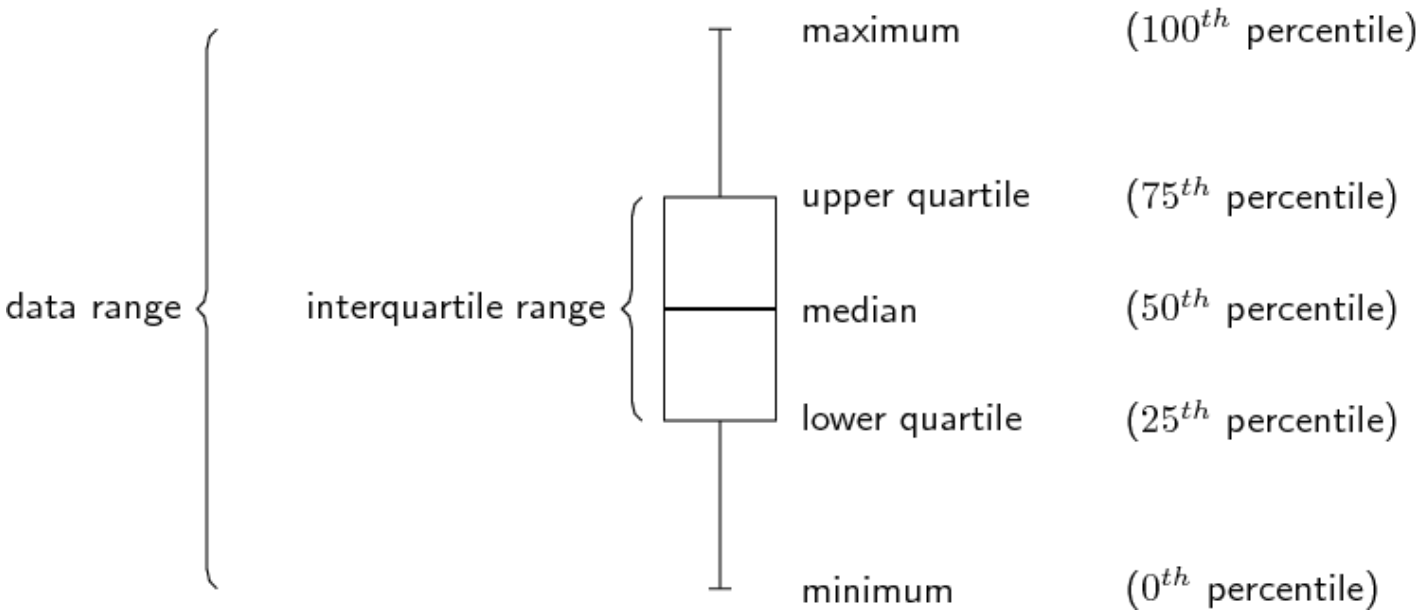
```
ggsave("H:/Jmburu/birth.pdf",plot=p1 ,width=8, height=6, unit="in")
```

Box plot

Box plot

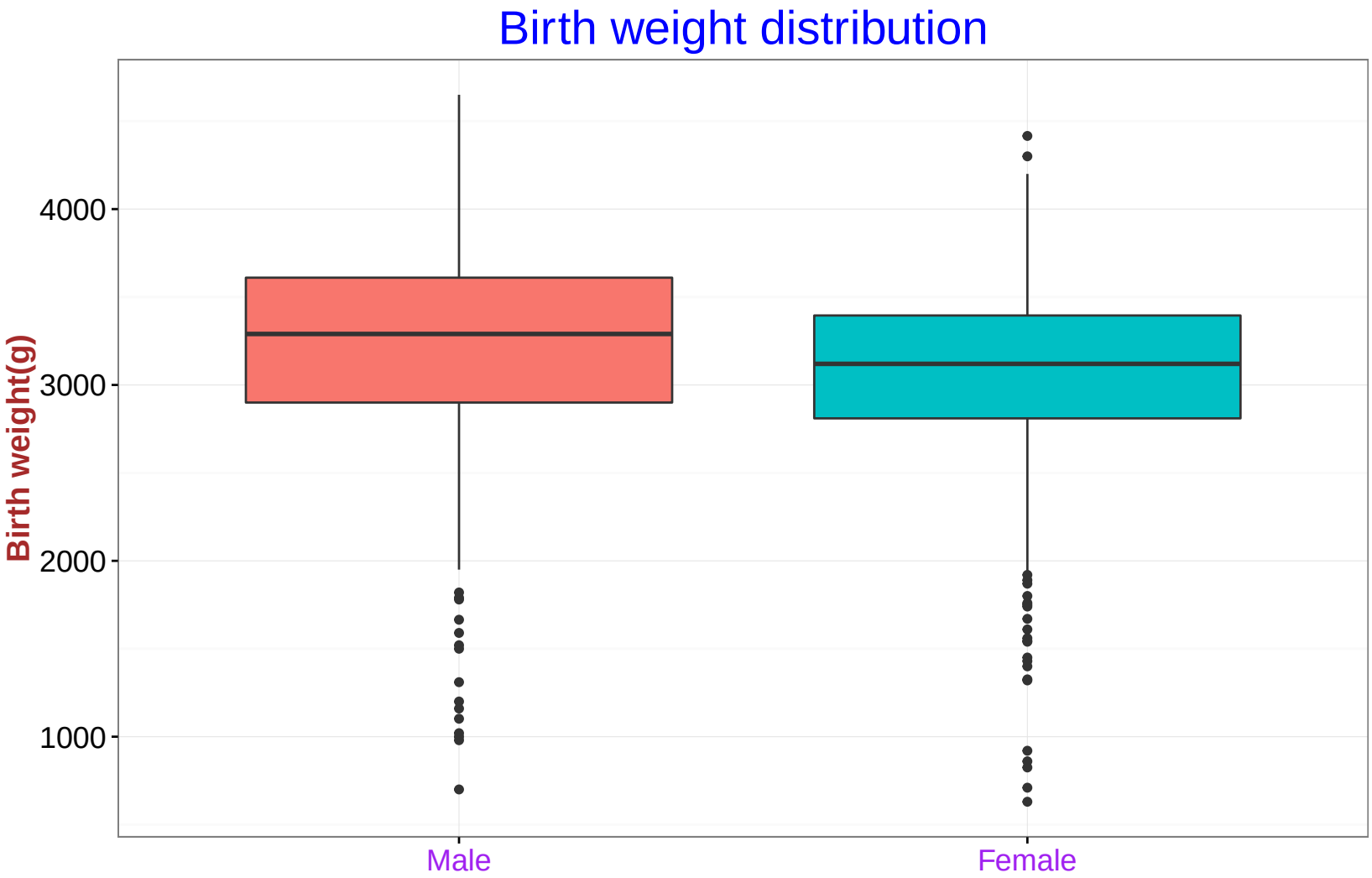
- Provides a standardized way of displaying the distribution of data.
- It attempts to provide a visual shape of the data distribution.
- This is based on some summary measures: min, 1^{st} quartile, median, 3^{rd} quartile, and max.
- Range, IQR, Outliers= $3^{*}IQR$ above 3^{rd} or below 1^{st} quartiles.

Box plot ...



–> ## Box plot ...

```
bx<-ggplot(birth, aes(x=sex, y=bweight, fill=sex)) +
    geom_boxplot() + guides(fill=FALSE)
bx<- bx + theme_bw() + labs(title="Birth weight distribution"
                            ,x="",y="Birth weight(g)")
bx<- bx + theme(title=element_text(color="blue", size=18),
  axis.title.y=element_text(size=15,face="bold",color="brown"),
```


Birth weight distribution
Birth weight(g)
4000
3000
2000
1000
Male
Female

Bar plot

Bar plot

- Provide a visual presentation of categorical data.
- Present grouped data with rectangular bars with lengths proportional to the values that they represent.
- Two types;
  - Grouped - presents bars clustered in groups
  - Stacked - shows bars divided into subparts to show cumulative effects.

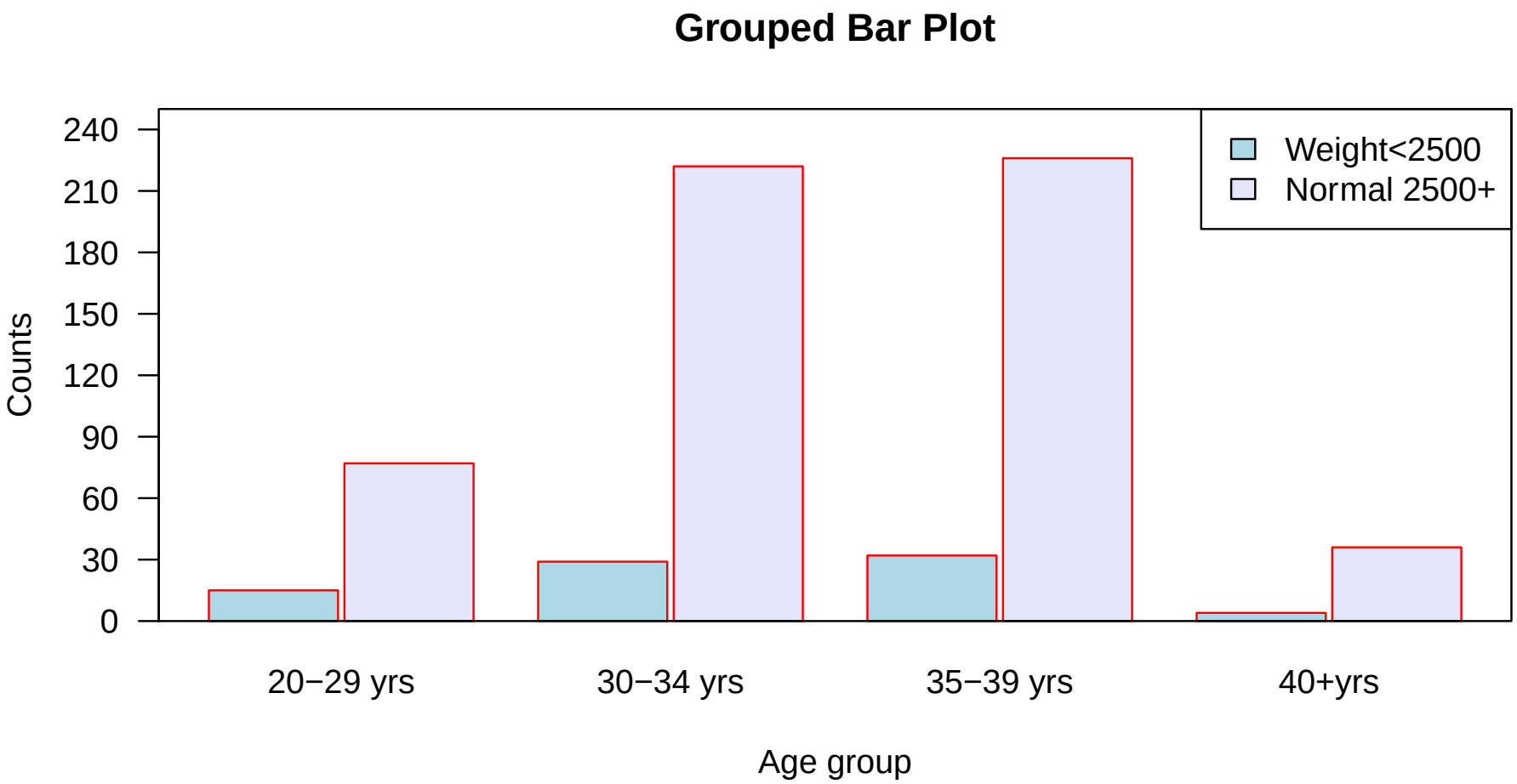
Bar plot . . .

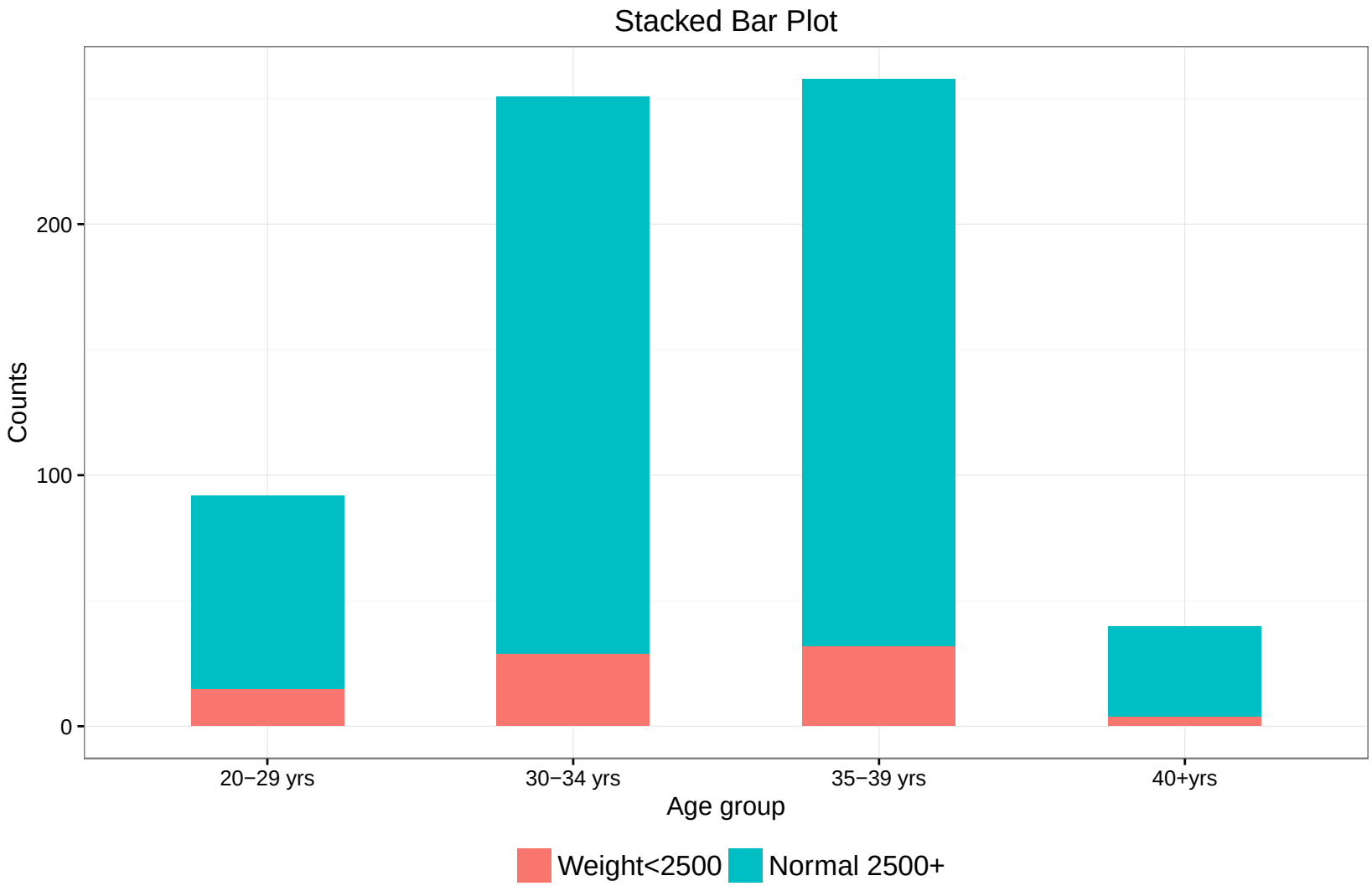
e.g. Cross tabulation of mother's age and birth weight.

	20-29 yrs	30-34 yrs	35-39 yrs	40+yrs
Weight <2500	15(0.16)	29(0.12)	32(0.12)	4(0.10)
Normal 2500+	77(0.84)	222(0.88)	226(0.88)	36 (0.90)
Total	92	251	258	40

Bar plot ... Grouped Bar Plot

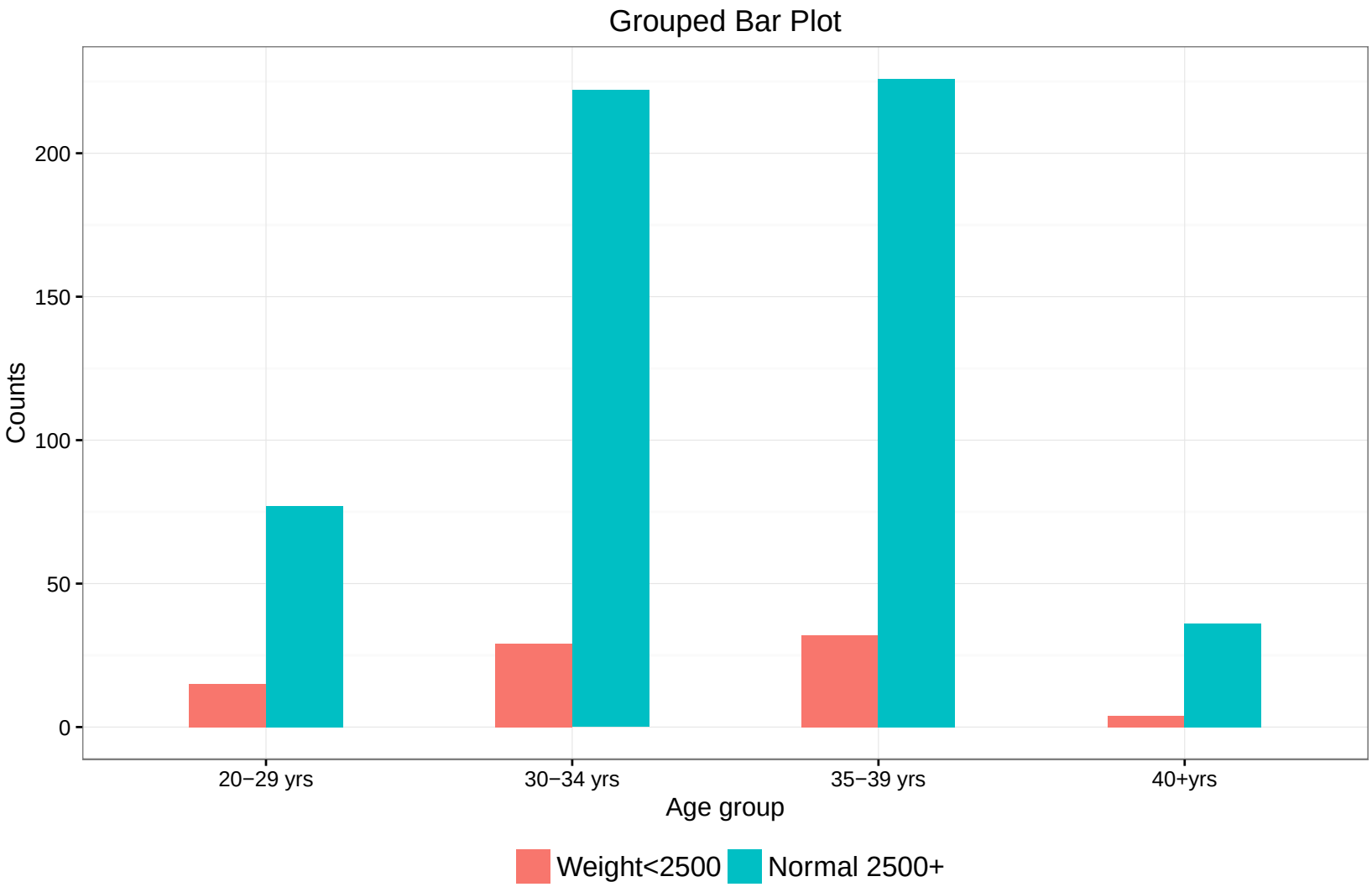
```
barplot(Count, beside = T, main ="Grouped Bar Plot",xlab="Age group",ylab ="Counts"
        , border ="red",yaxt='n',col=c("lightblue","lavender"),ylim=c(0,250),
        legend=rownames(Count),args.legend =list(x ="topright"),space=c(0.05,0.5))
axis(2, at = seq(0, 250, by = 30), las = 1);box()
```




Stacked Bar Plot
Counts
200
100
0
20−29 yrs
30−34 yrs
35−39 yrs
40+yrs
Age group
Weight<2500
Normal 2500+

Bar plot ... Grouped Bar Plot

```
b<-ggplot(data=Count, aes(x=agegrp,y=count,fill=lbw))
b<- b + geom_bar(stat="identity",width=.5,position ="dodge") +
  theme_bw()
b<- b + labs(title="Grouped Bar Plot",x="Age group",y="Counts")
b<- b + theme(legend.key=element_blank(),
           legend.title=element_blank(),
           legend.text=element_text(size = rel(1.1)),
           legend.direction="horizontal",legend.position="bottom")
```


Grouped Bar Plot
Counts
200
150
100
50
0
20−29 yrs
30−34 yrs
35−39 yrs
40+yrs
Age group
Weight<2500
Normal 2500+

Links

Cookbook for R

Help topics

Thank You