EDA gh.v2

Barbara E. Mottey

11/8/2020

Load data

```
gh_data <- readRDS("named_datagh.rds")
```

##birth weight and fuel

```
gh.data <- gh_data
attach(gh.data)
summary(c_weight) # 28,699 NAs Min is 500 and max is 9000
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##     500    2800    3100    3205    3500    9000   28699
```

```
table(fuel_bin) # 801 NAs
```

```
## fuel_bin
##     0     1
##  2868 16700
```

```
#(k<-gh.data %>% group_by(year) %>% summarise(observations= n()))
filter(gh_data, c_weight> 5500) %>% summarise(n=n())  # 34 obs are greater than 5500
```

```
##    n
## 1 34
```

```
#bw 5884, 1971 not weighed (coded 9996), dont know 543 NA 3
#2008 2903 2290 613 weighed
## read in, filter by bw<=5500 and compare with what the FRQ file has.
```

refine data

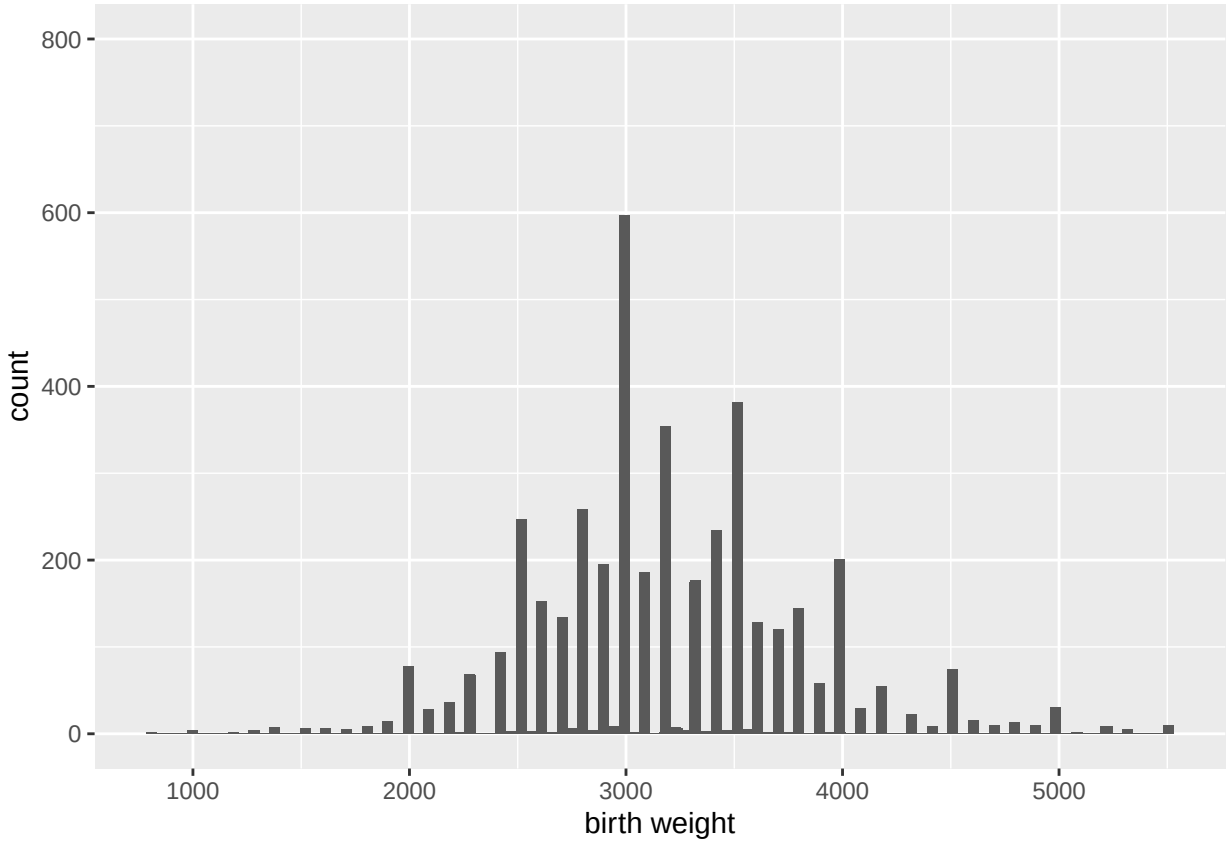
```
# remove bw above 5500 and bw NA's
gh.data <- gh.data %>% filter(c_weight <= 5500) #28746 total removed

#filter data from year 2000

filter(gh.data, year_cmc> 2000) %>% summarise(n=n()) # total of 4299 data
```

```
##      n
## 1 4299
```

```
gh.data <- gh.data %>% filter(year_cmc > 2000)

#save data
saveRDS(gh.data, "model_datagh.rds")
```

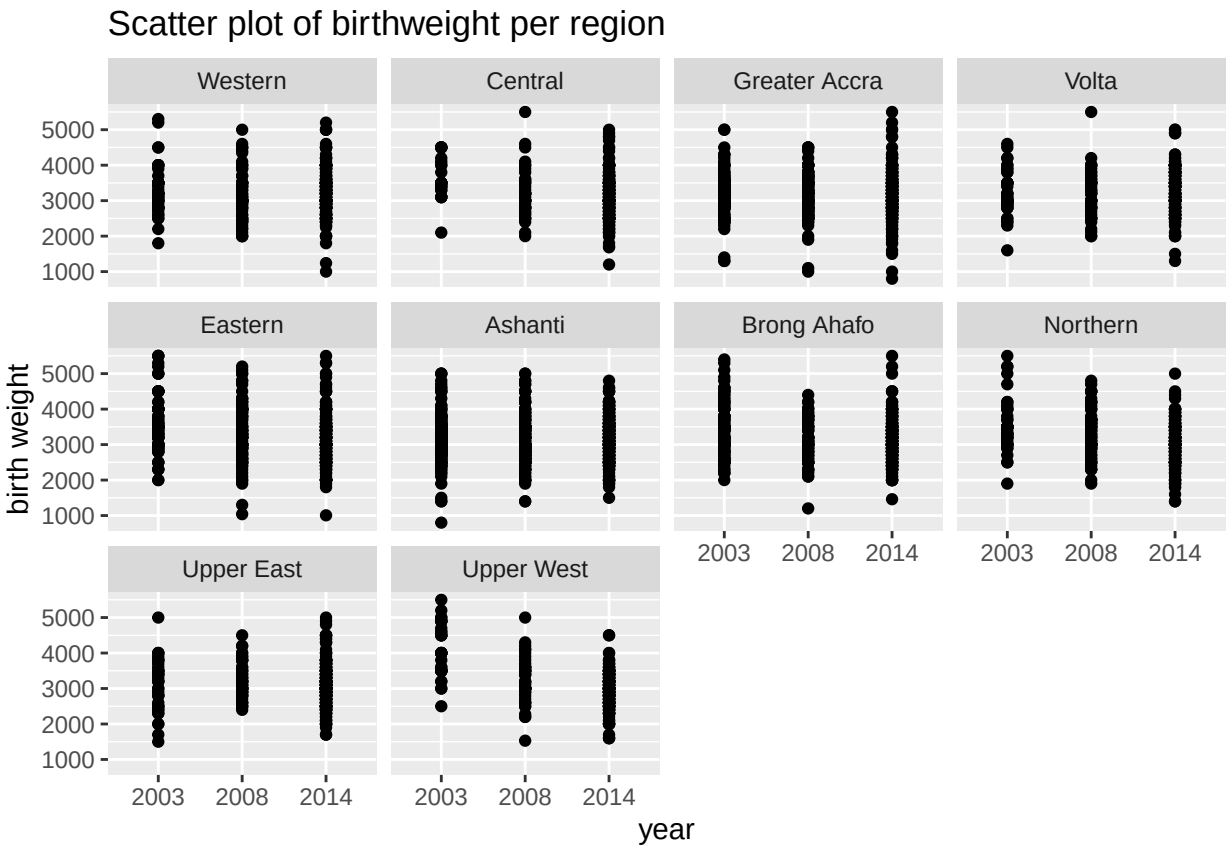
birth weight distribution

```
ggplot(gh.data, aes(c_weight))+geom_histogram(binwidth = 10.5, alpha = 1.0)+ ylim(0,800)+stat_bin(bins=
```

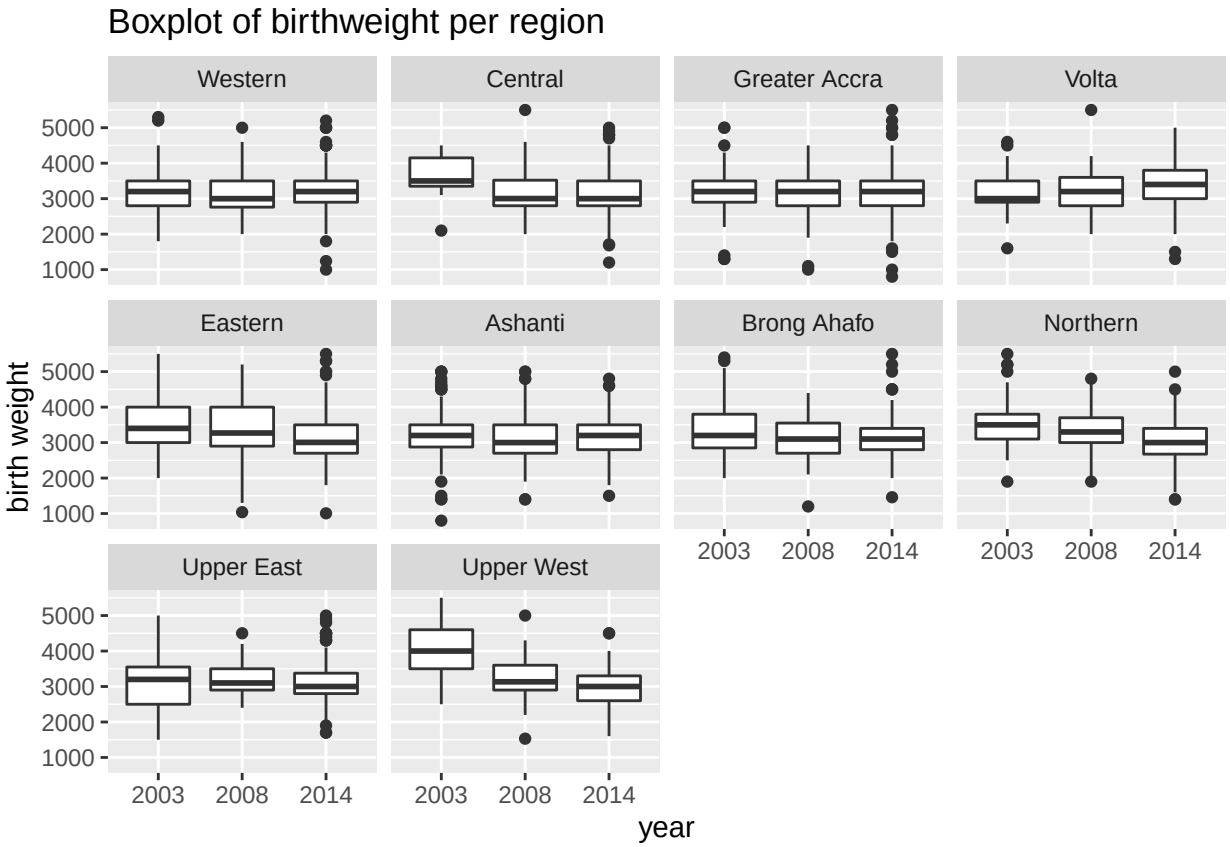


birthweight vs year per region

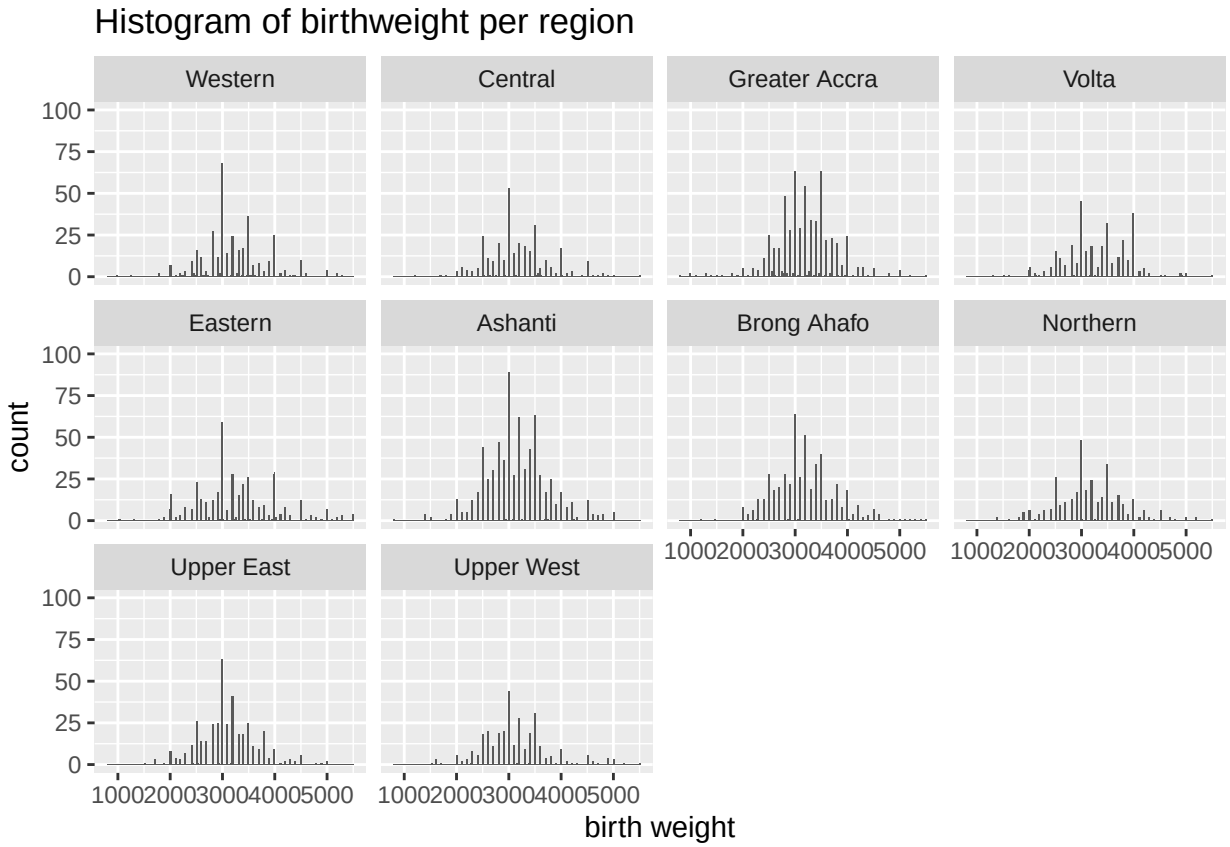
```
ggplot(gh.data)+ aes(as.factor(year_cmc), c_weight)+geom_point()+ facet_wrap(~region, labeller=(as_label
```




```
ggplot(gh.data)+ aes(as.factor(year_cmc), c_weight)+geom_boxplot()+ facet_wrap(~region, labeller=(as_la
```

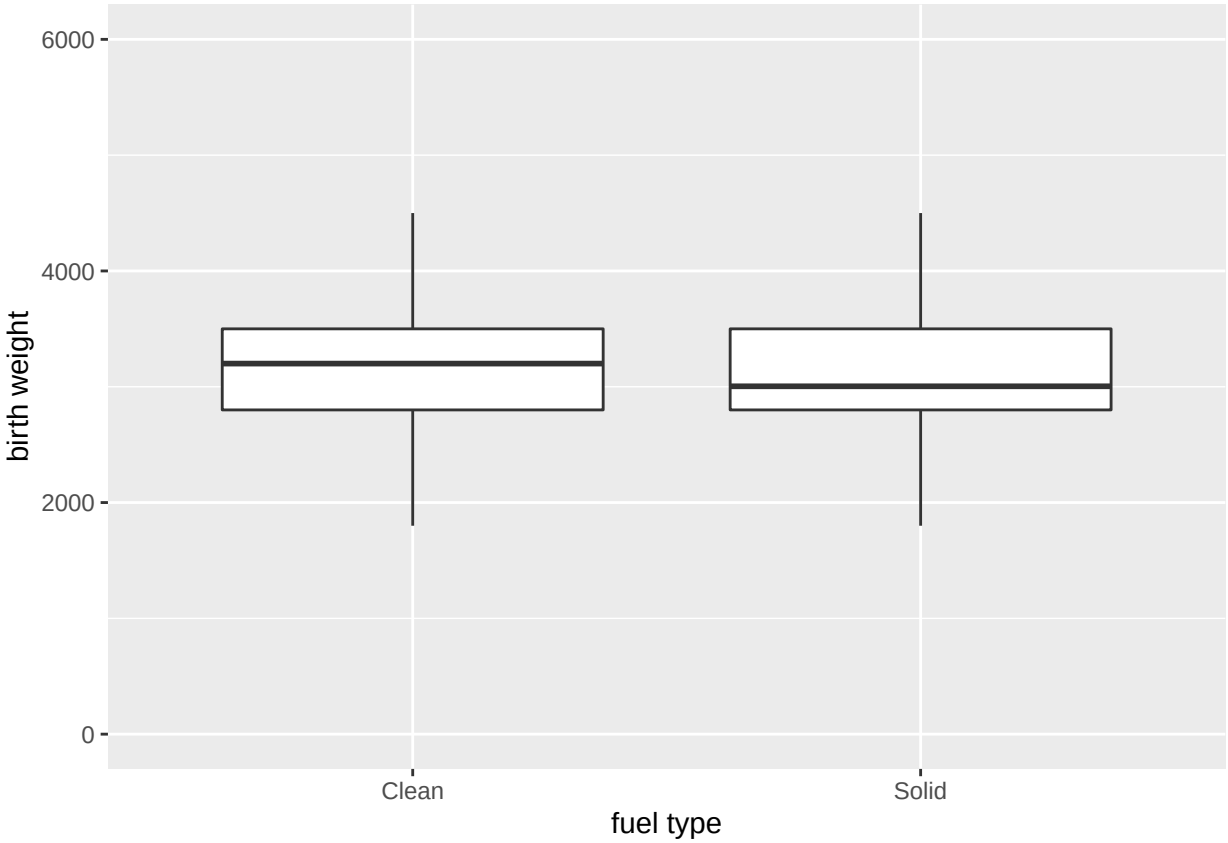



```
ggplot(gh.data, aes(c_weight))+geom_histogram(binwidth = 10.5, alpha = 1.0)+ ylim(0,100)+stat_bin(bins=
```



boxplots for fuel types

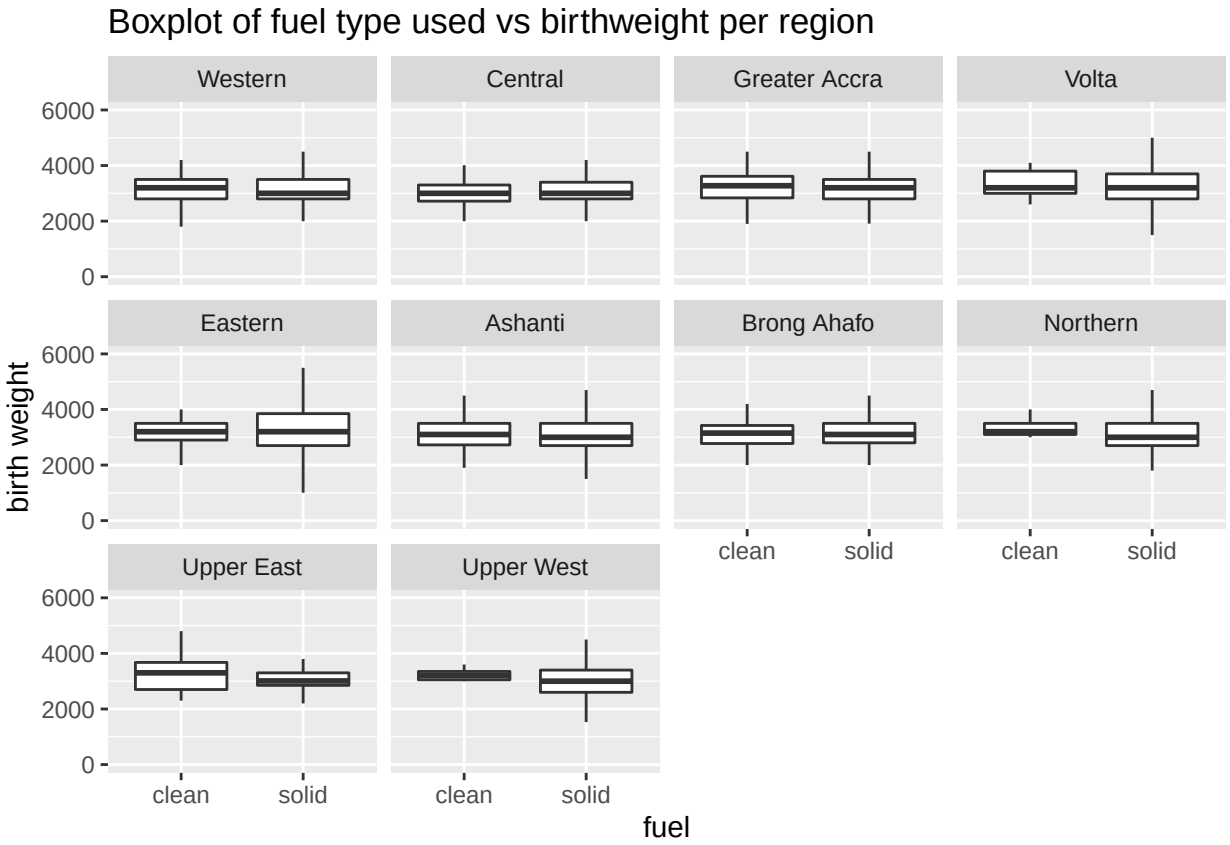
```
gh.data$fuel_bin[gh.data$fuel_bin == 2]<- NA
gh.data$fuel_bin <- as.factor(gh.data$fuel_bin)

ggplot(na.omit(gh.data), aes(x= fuel_bin, y= c_weight))+geom_boxplot(outlier.shape = NA)+ ylim(0,6000) +
  scale_x_discrete(labels = c("0" = "Clean", "1"  = "Solid"))
```



Generally, people who use solid fuel tend to have babies with lower birthweights comparatively.

##per region

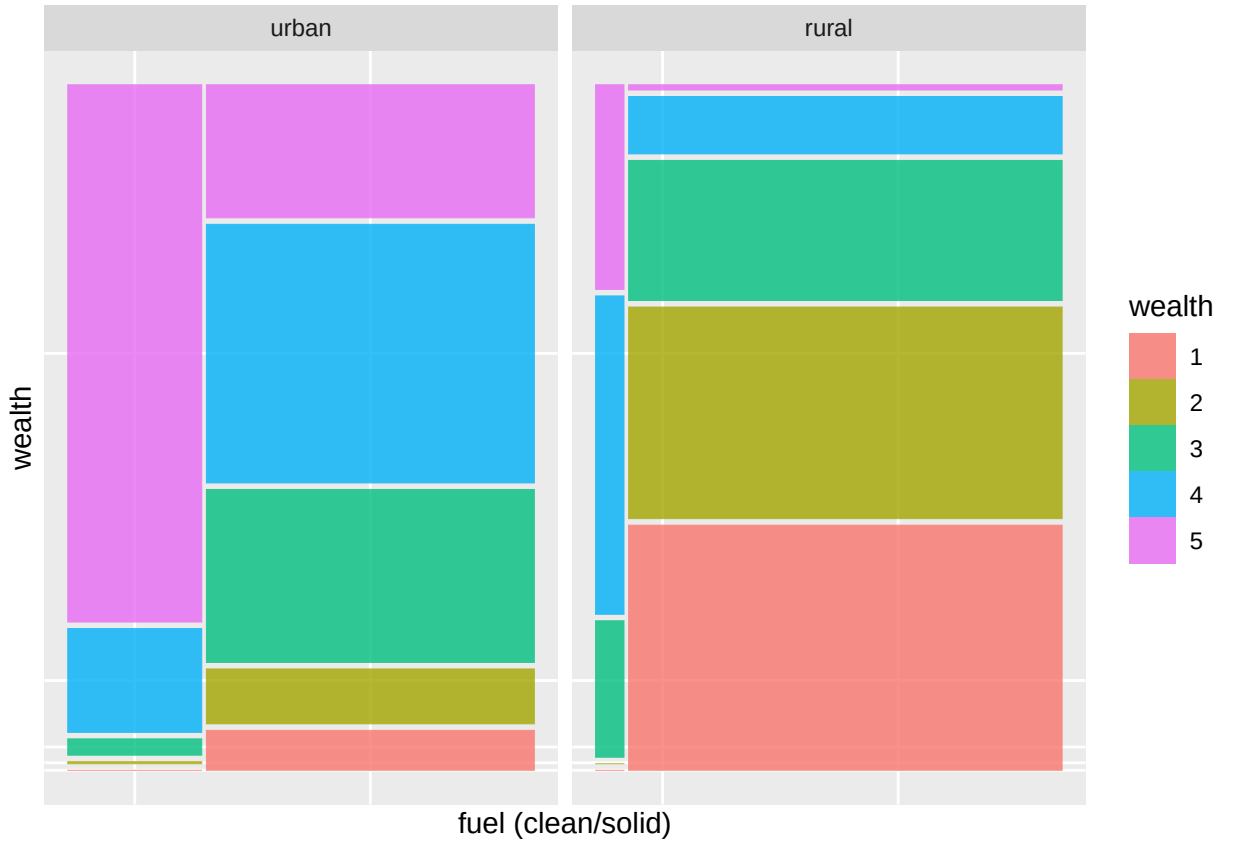
```
ggplot(na.omit(gh.data), aes(x= fuel_bin, y= c_weight))+geom_boxplot(outlier.shape = NA)+ ylim(0,6000)+
```



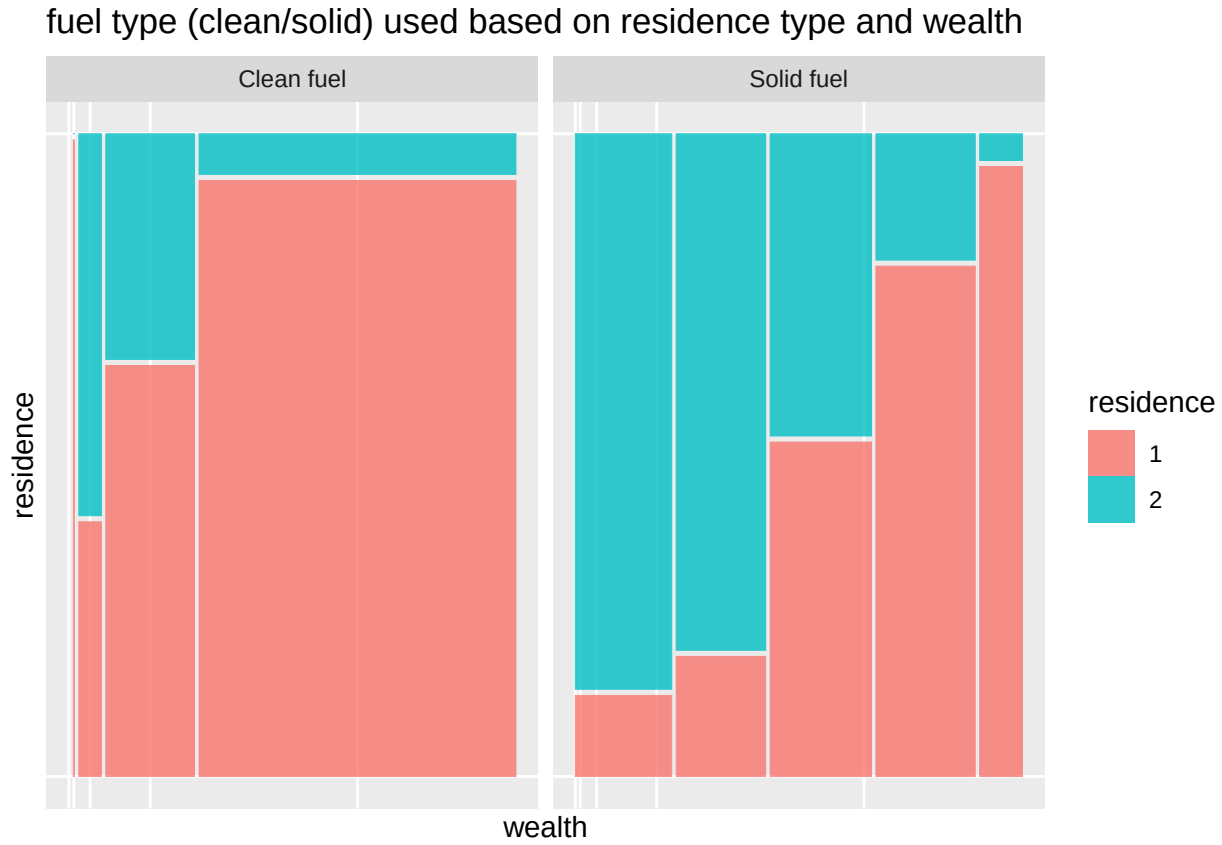
Stratified by region, there seems to be no difference in birthweights for babies with mothers that use solid or clean fuel.

###fuel type vs wealth

```
gh.data$residence<- as.factor(gh.data$residence)
gh.data$wealth <- as.factor(gh.data$wealth)

ggplot(na.omit(gh.data)) + geom_mosaic(aes(product(wealth, fuel_bin), fill = wealth)) + facet_wrap(~resi
```




```
ggplot(na.omit(gh.data)) + geom_mosaic(aes(product(residence, wealth), fill = residence)) +
  facet_wrap(~fuel_bin, labeller=(as_labeller(c("0" = "Clean fuel", "1"  = "Solid fuel")))) +labs(x= "w
```

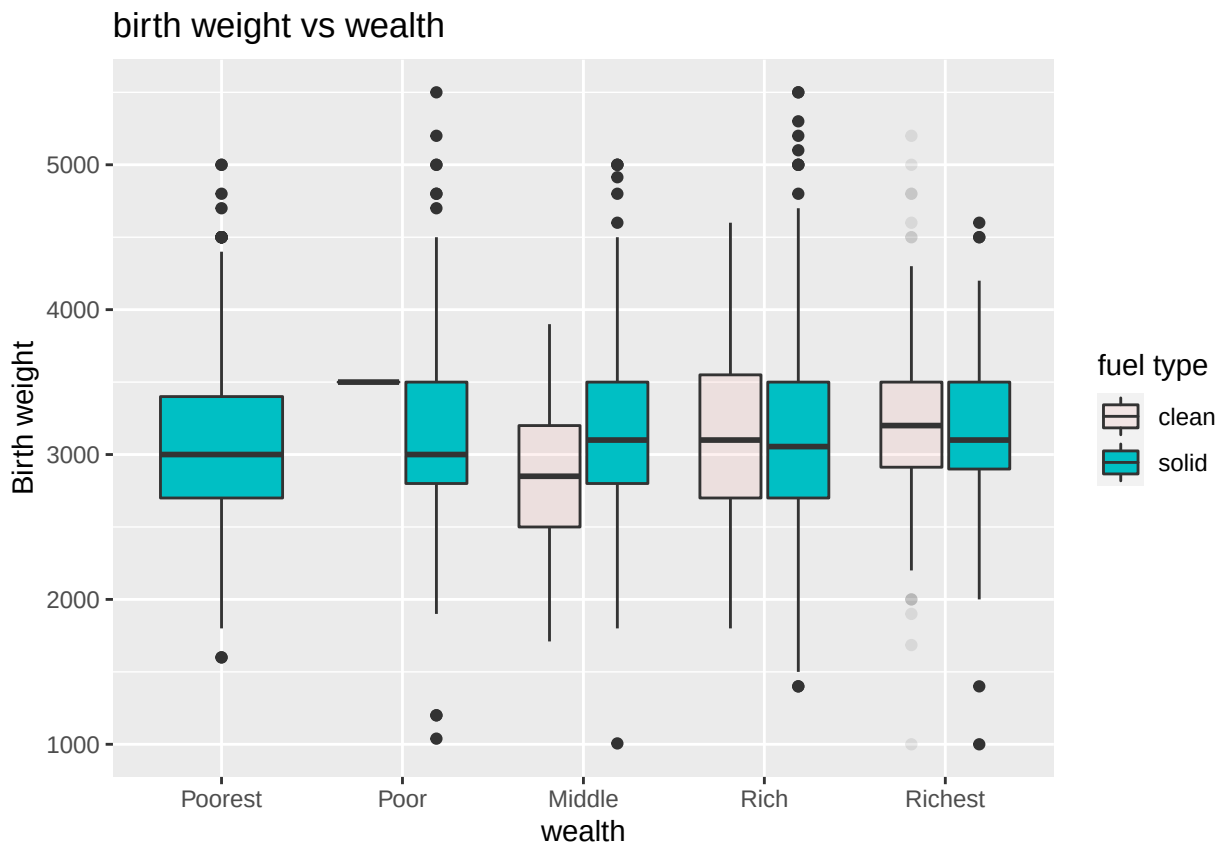


#Facet by Residential Among urban residents, a higher proportion of those who use clean fuels are the rich people. About the 80% of those who use solid fuel are at least average people in terms of wealth. However among the rural residents, a larger proportion (about 90%) use solid fuel. Most of those who use sold fuel in the rural setting are poor where as those who use clean fuel (about 98%) have at least average wealth.

#Facet by Fuel The higher the wealth index, the more likely the residence is urban for the clean fuel users.

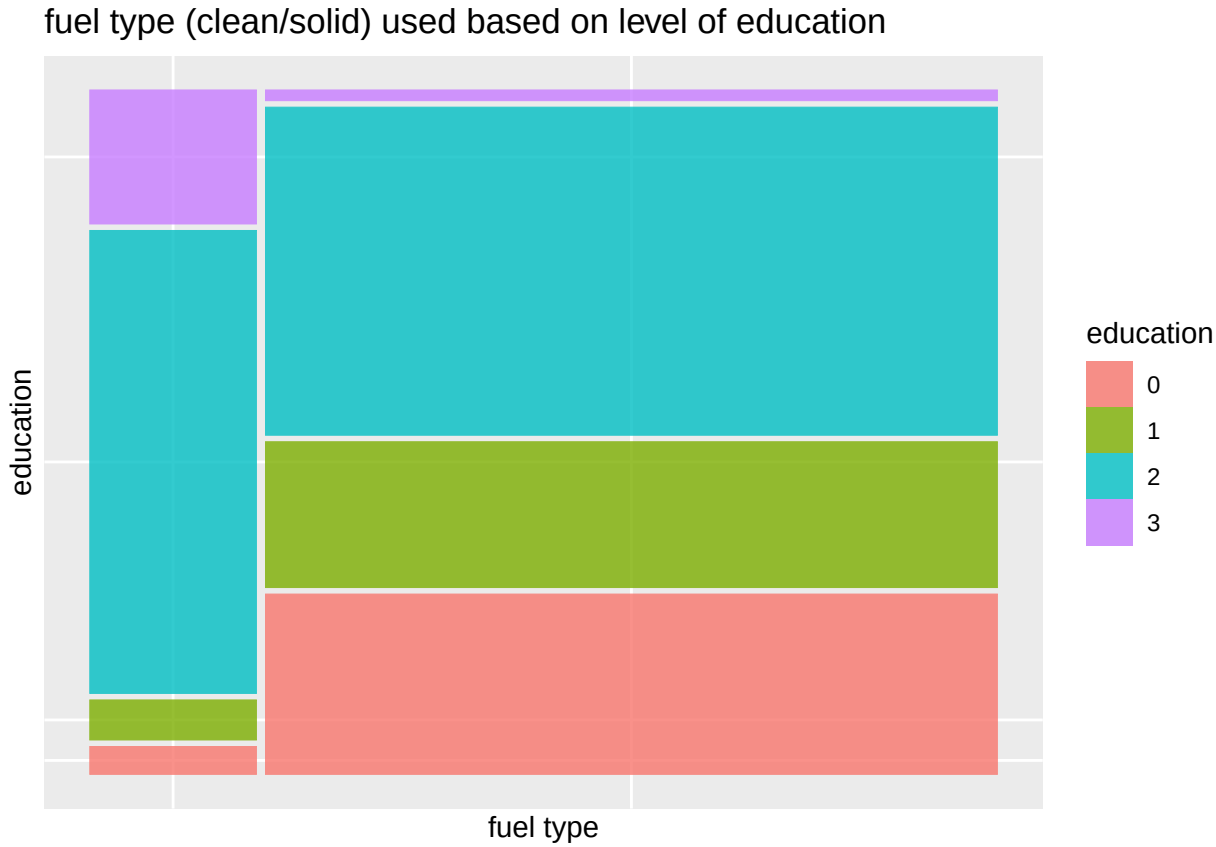
Fuel vs wealth

```
ggplot(na.omit(gh.data))+ geom_boxplot(aes(wealth, c_weight, fill = fuel_bin, alpha = fuel_bin))+
  labs(x = "wealth", y = "Birth weight", title = "birth weight vs wealth")+
  scale_x_discrete(labels = c("1"= "Poorest", "2"= "Poor", "3" = "Middle", "4" = "Rich", "5" = "Richest
  scale_alpha_ordinal(name = "fuel type", labels = c("clean", "solid"))+
  scale_fill_discrete(name = "fuel type", labels = c("clean", "solid"))
```



##Fuel vs education

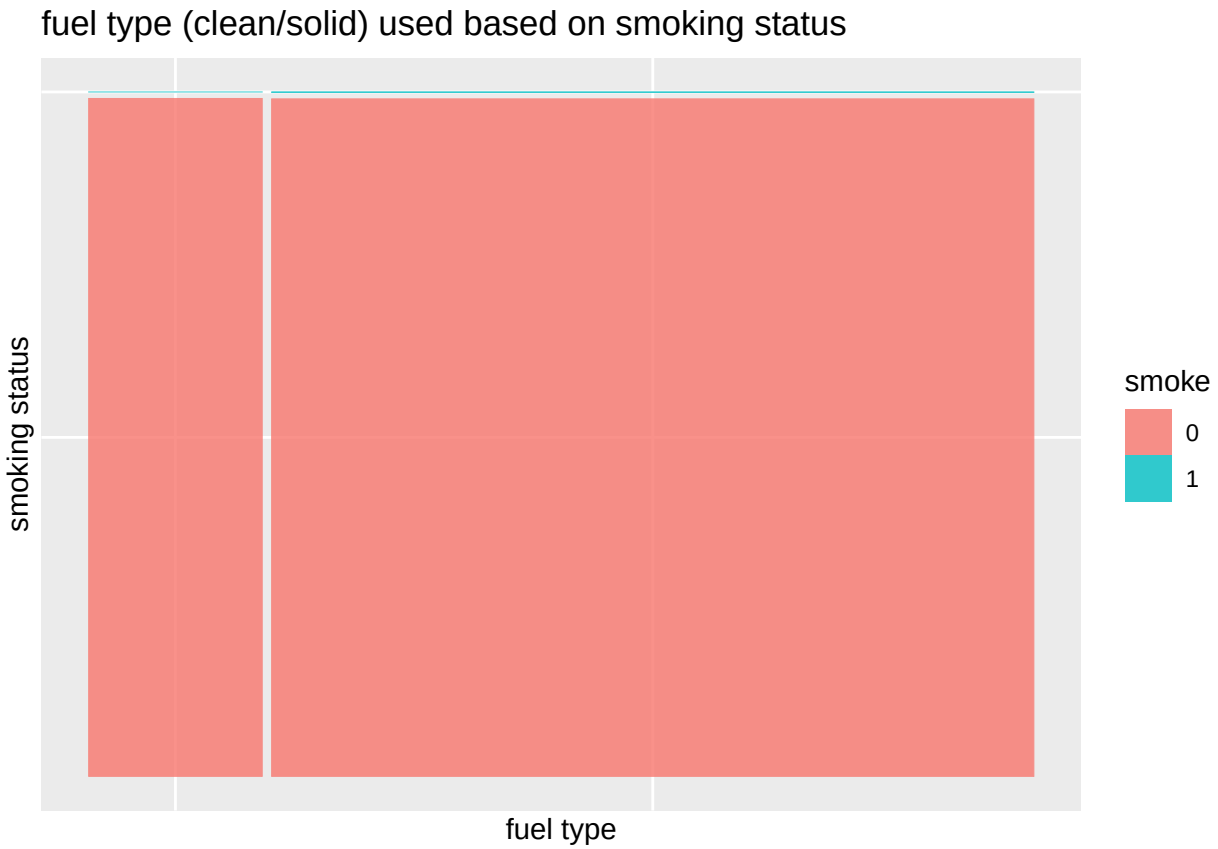
```
gh.data$education <- as.factor(gh.data$education)

ggplot(na.omit(gh.data)) + geom_mosaic(aes(product(education, fuel_bin), fill = education)) + labs(x= "
```



Most of the women that use clean fuel have a secondary education; same applies to those who use solid fuel. About 50% of those who use solid fuel have at most primary education.

fuel vs smoking

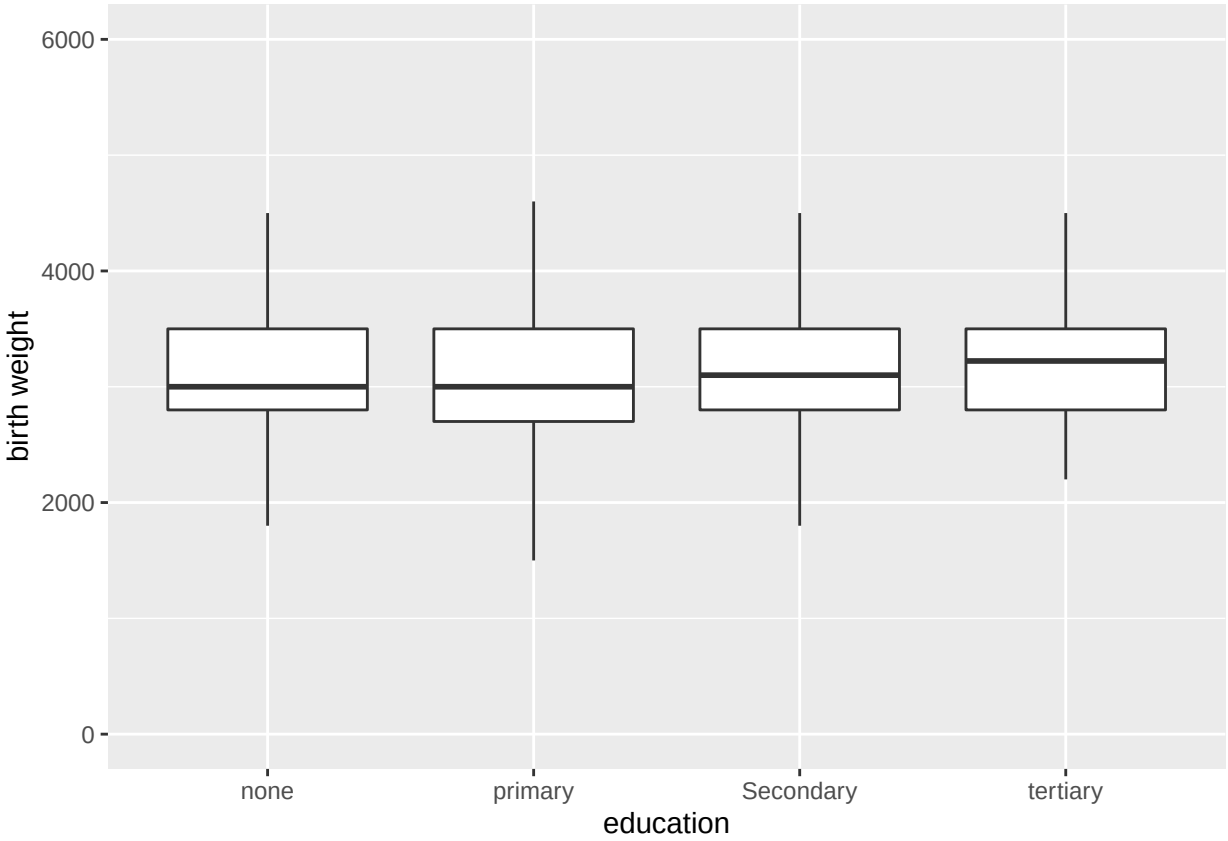
```
gh.data$smoke <- as.factor(gh.data$smoke)

ggplot(na.omit(gh.data)) + geom_mosaic(aes(product(smoke, fuel_bin), fill = smoke)) + labs(x= "fuel typ
```



Smoking status would not be a covariate. Most women do not smoke.

##education vs birth weight

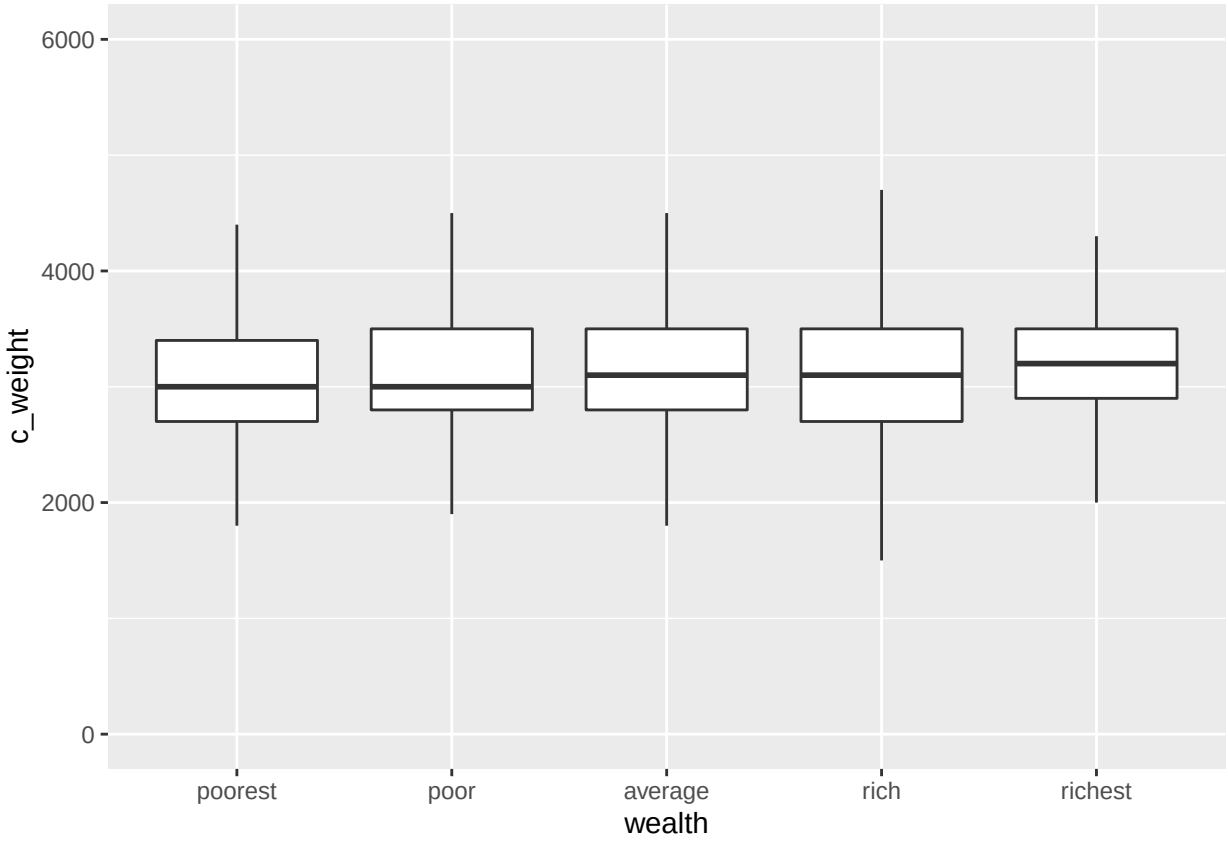
```
ggplot(na.omit(gh.data), aes(x= education, y= c_weight))+geom_boxplot(outlier.shape = NA)+ ylim(0,6000)
```



Those with no education or primary education seem to have babies with a lower mean weight.

##birth weight vs wealth

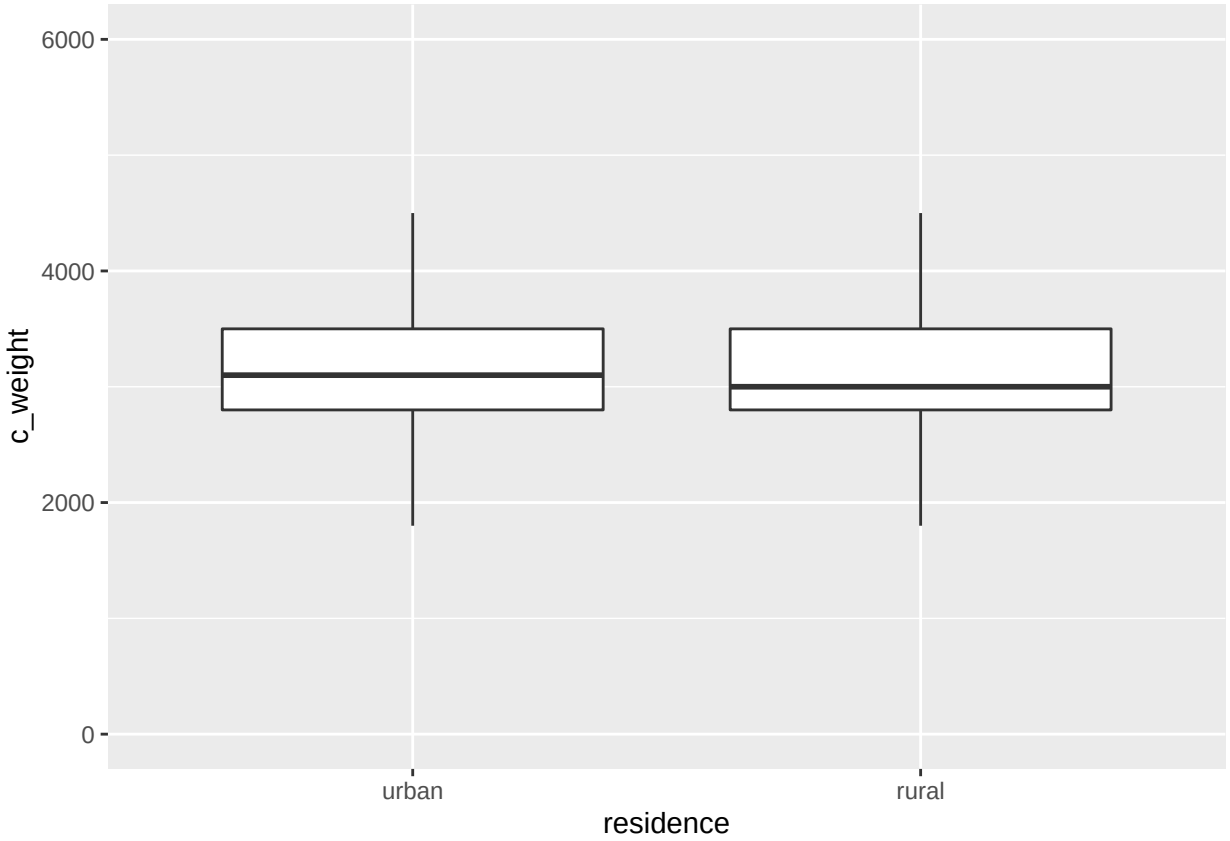
```
ggplot(na.omit(gh.data), aes(x= wealth, y= c_weight))+geom_boxplot(outlier.shape = NA)+ ylim(0,6000)+sc
```



The mean birth weight for each category seem to be lower for those who are poorest or poor.

birth weight vs residence

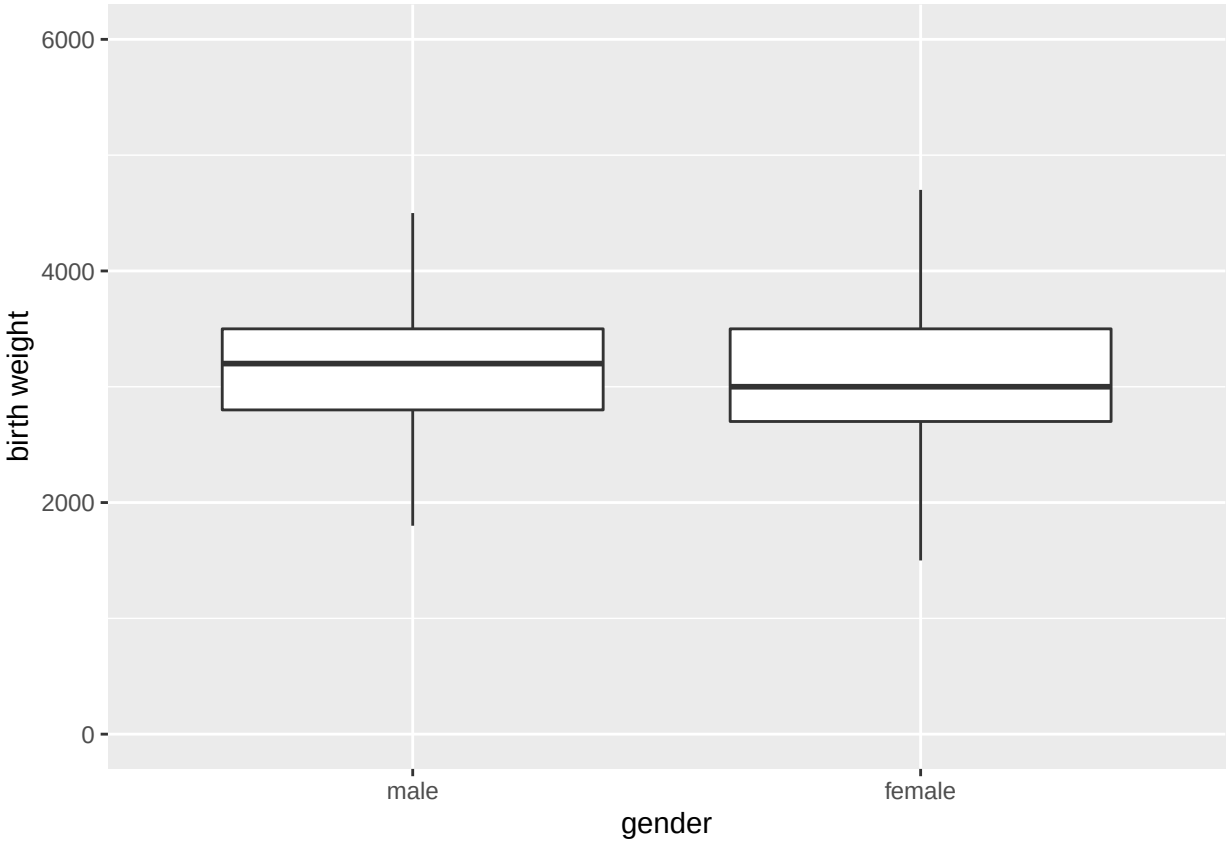
```
ggplot(na.omit(gh.data), aes(x= residence, y= c_weight))+geom_boxplot(outlier.shape = NA)+ ylim(0,6000)
```



Average Birth weights for those in rural settings seem to be quite lower than those in the urban setting.

birth weight vs gender

```
gh.data$gender <- as.factor(gh.data$gender)
ggplot(na.omit(gh.data), aes(x= gender, y= c_weight))+geom_boxplot(outlier.shape = NA)+ ylim(0,6000)+ la
```



boys seem to have a higher average birth weight than girls