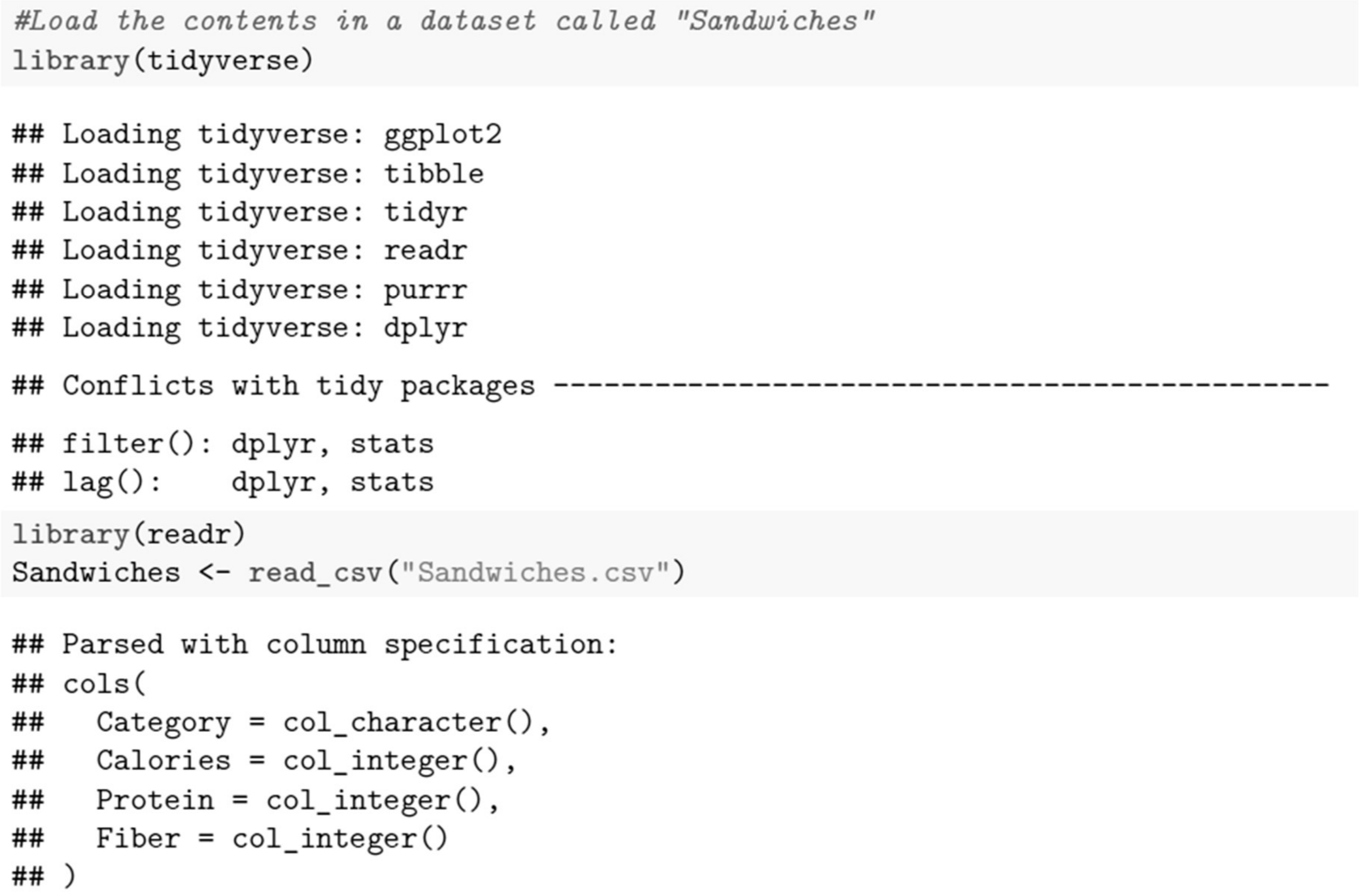
Data Summarizing & Cleaning

Loading packages & importing datasets



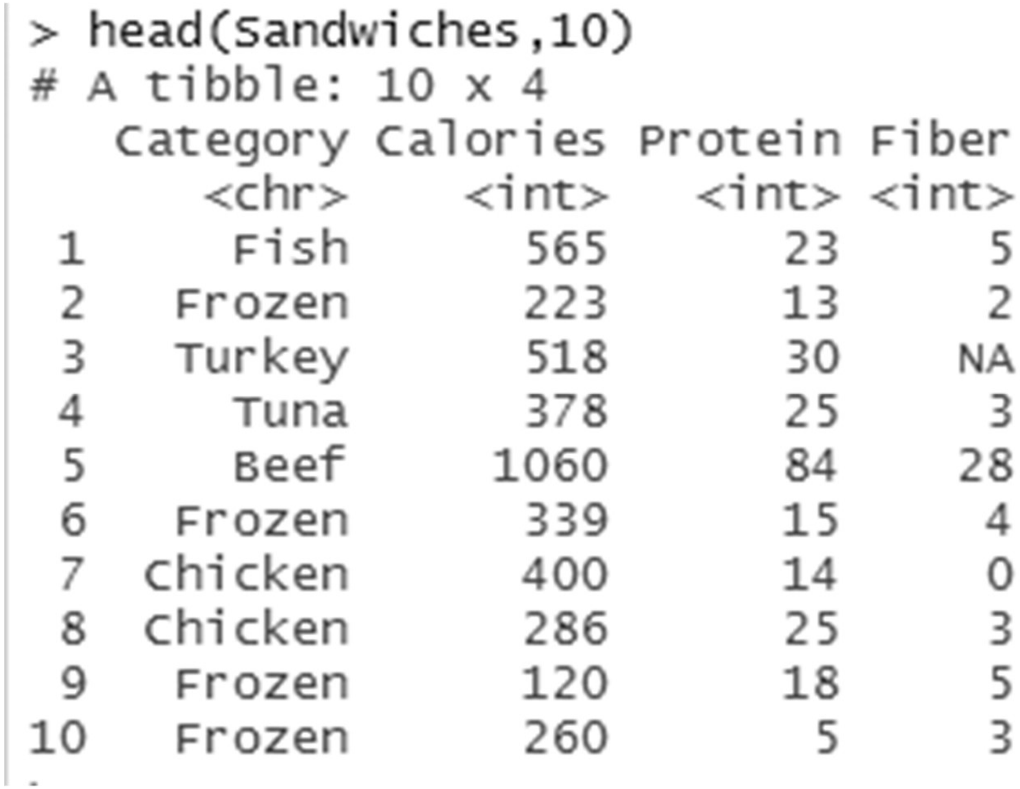
```
#Load the contents in a dataset called "Sandwiches"
library(tidyverse)
```

```
## Loading tidyverse: ggplot2
## Loading tidyverse: tibble
## Loading tidyverse: tidyr
## Loading tidyverse: readr
## Loading tidyverse: purrr
## Loading tidyverse: dplyr

## Conflicts with tidy packages ----------------------------------------------

## filter(): dplyr, stats
## lag():    dplyr, stats
```

```
library(readr)
Sandwiches <- read_csv("Sandwiches.csv")
```

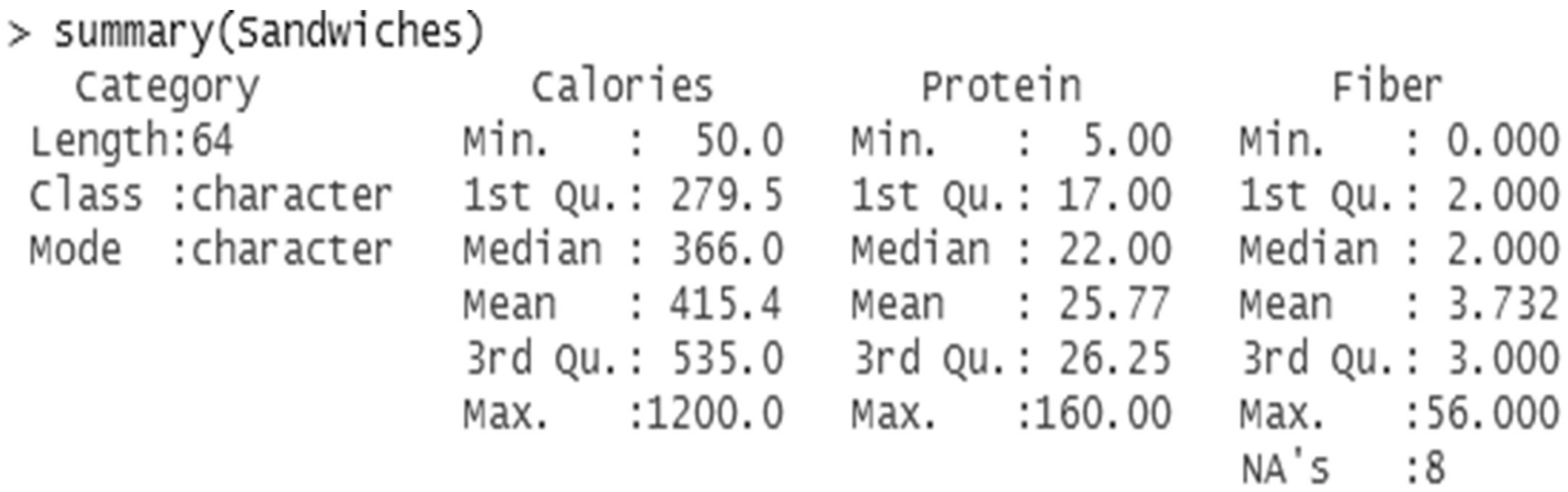
```
## Parsed with column specification:
## cols(
##   Category = col_character(),
##   Calories = col_integer(),
##   Protein = col_integer(),
##   Fiber = col_integer()
## )
```

View the contents of table and examine the structure



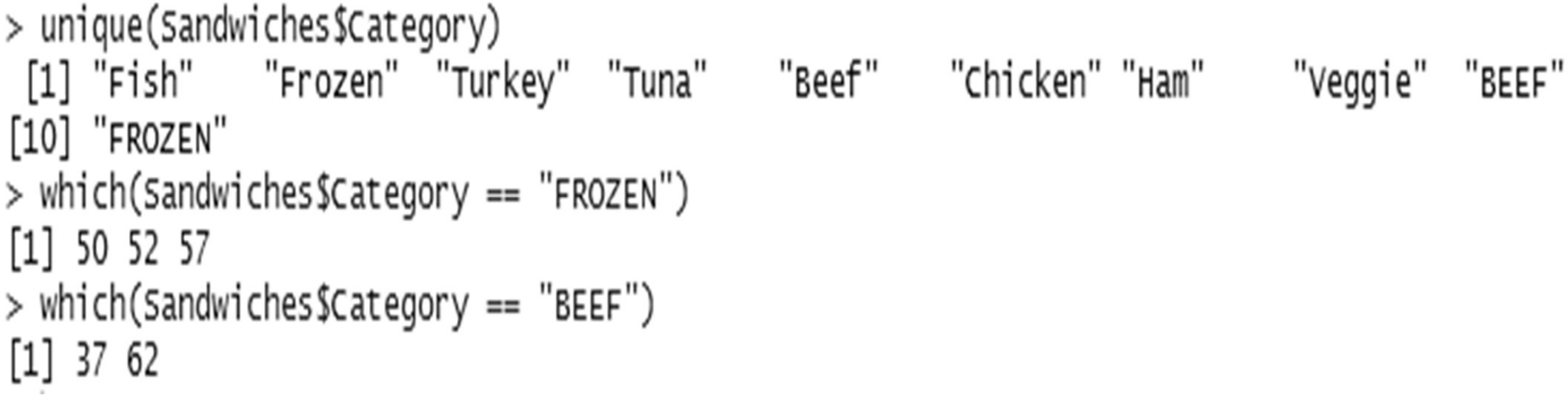
```
> head(Sandwiches,10)
# A tibble: 10 x 4
   Category Calories Protein Fiber
      <chr>    <int>   <int> <int>
 1     Fish      565      23     5
 2   Frozen      223      13     2
 3   Turkey      518      30    NA
 4     Tuna      378      25     3
 5     Beef     1060      84    28
 6   Frozen      339      15     4
 7  Chicken      400      14     0
 8  Chicken      286      25     3
 9   Frozen      120      18     5
10   Frozen      260       5     3
```

Summarize the dataset



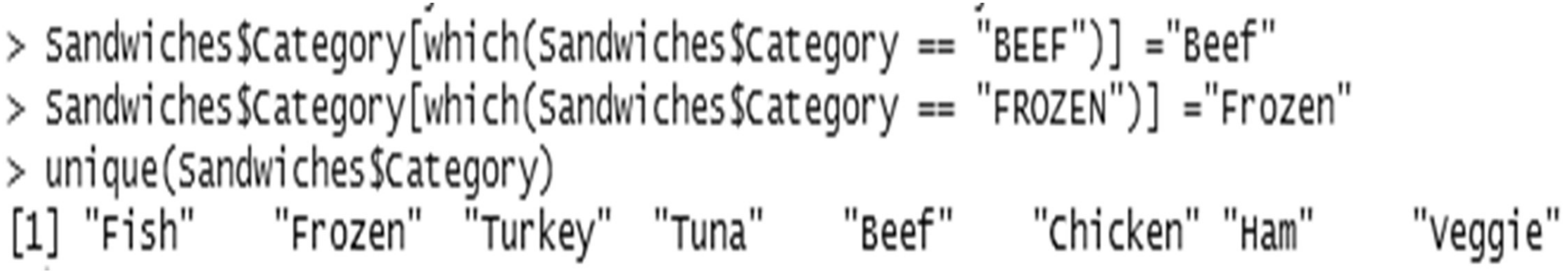
```
> summary(Sandwiches)
   Category            Calories          Protein            Fiber       
 Length:64          Min.   :  50.0   Min.   :  5.00   Min.   : 0.000  
 Class :character   1st Qu.: 279.5   1st Qu.: 17.00   1st Qu.: 2.000  
 Mode  :character   Median : 366.0   Median : 22.00   Median : 2.000  
                    Mean   : 415.4   Mean   : 25.77   Mean   : 3.732  
                    3rd Qu.: 535.0   3rd Qu.: 26.25   3rd Qu.: 3.000  
                    Max.   :1200.0   Max.   :160.00   Max.   :56.000  
                                                      NA's   :8       
```

Identify data entry inconsistencies/errors in the variable



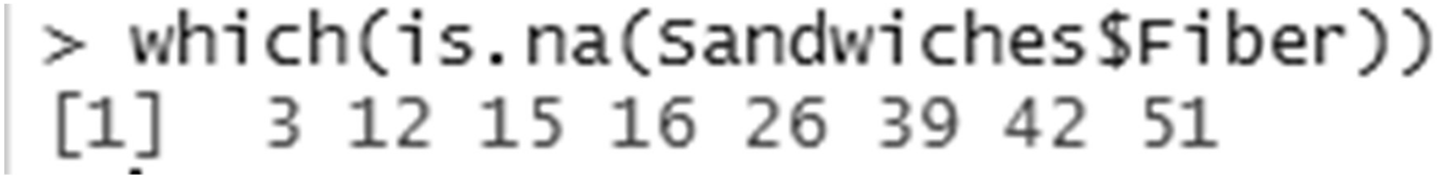
```
> unique(Sandwiches$Category)
 [1] "Fish"    "Frozen"  "Turkey"  "Tuna"    "Beef"    "Chicken" "Ham"     "Veggie"  "BEEF"
[10] "FROZEN"
> which(Sandwiches$Category == "FROZEN")
[1] 50 52 57
> which(Sandwiches$Category == "BEEF")
[1] 37 62
```

Clean the data entry errors to ensure consistency of text



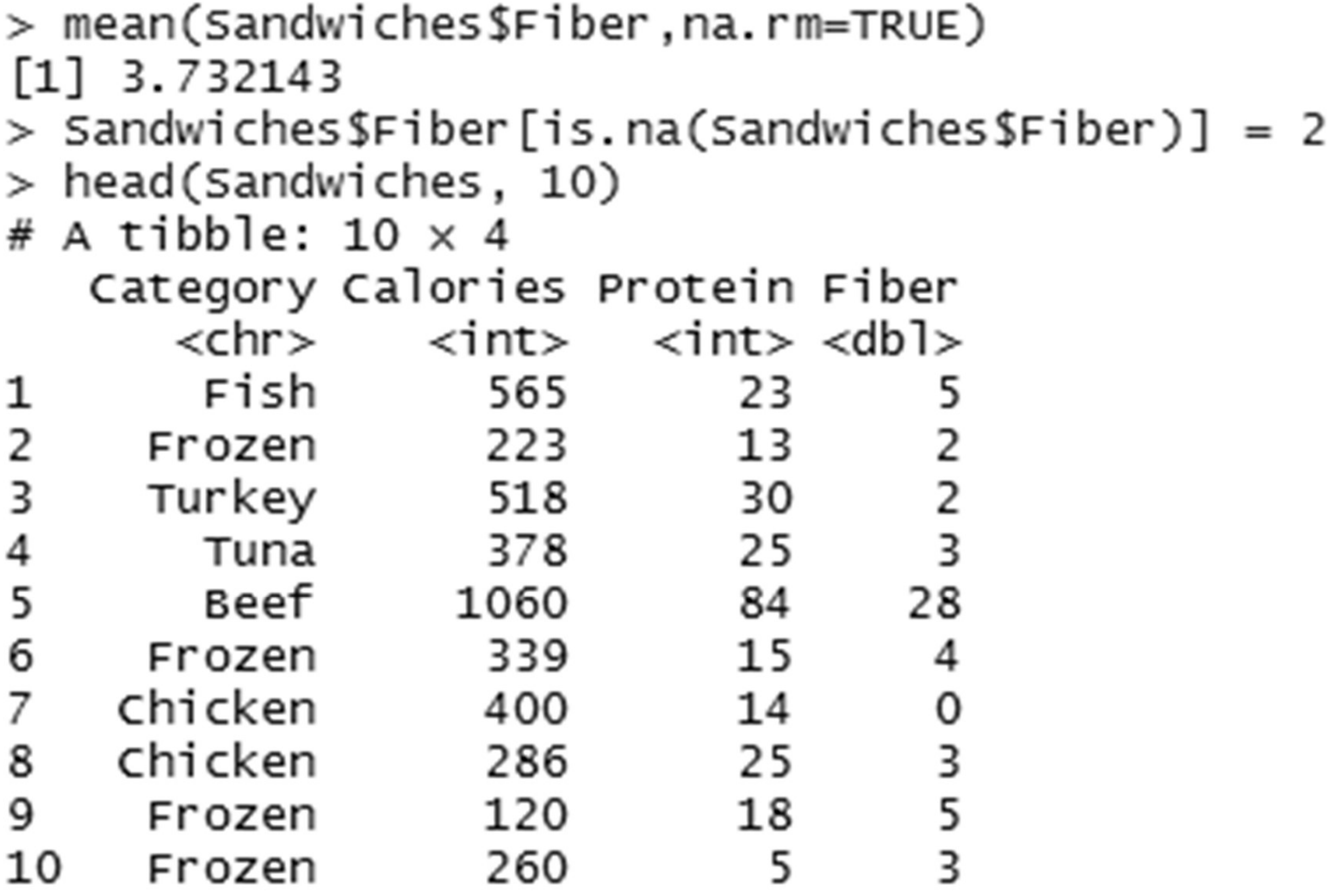
```
> Sandwiches$Category[which(Sandwiches$Category == "BEEF")] ="Beef"
> Sandwiches$Category[which(Sandwiches$Category == "FROZEN")] ="Frozen"
> unique(Sandwiches$Category)
[1] "Fish"    "Frozen"  "Turkey"  "Tuna"    "Beef"    "Chicken" "Ham"     "Veggie"
```

Identify location of missing values in the Fiber column



```
> which(is.na(Sandwiches$Fiber))
[1]  3 12 15 16 26 39 42 51
```

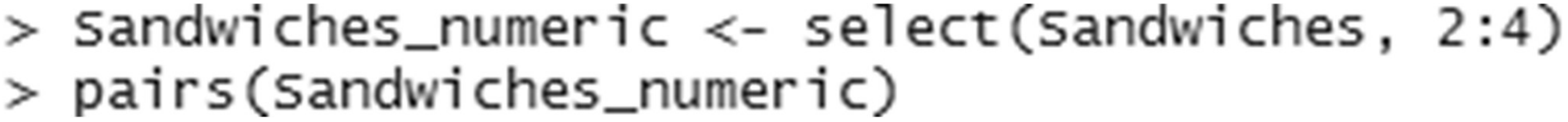
Compute the mean of Fiber while retaining missing values & impute Fiber = 2



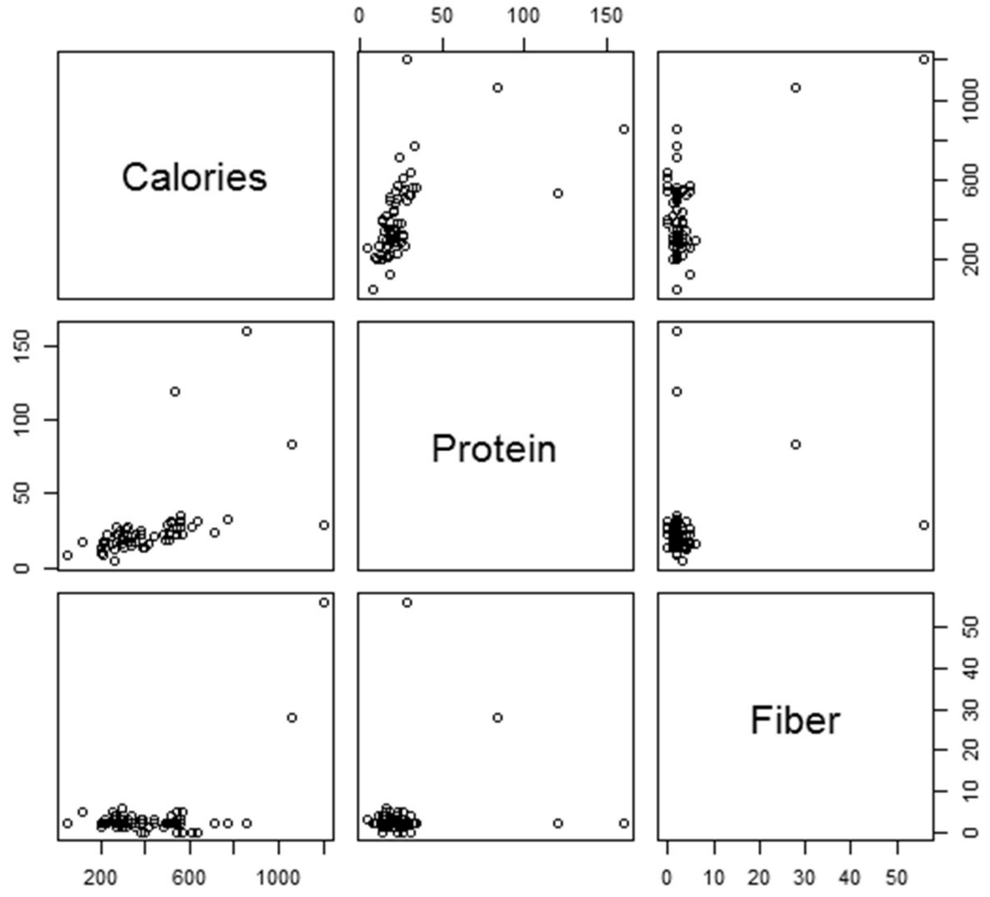
```
> mean(Sandwiches$Fiber,na.rm=TRUE)
[1] 3.732143
> Sandwiches$Fiber[is.na(Sandwiches$Fiber)] = 2
> head(Sandwiches, 10)
# A tibble: 10 x 4
   Category Calories Protein Fiber
      <chr>    <int>   <int> <dbl>
1      Fish      565      23     5
2    Frozen      223      13     2
3    Turkey      518      30     2
4      Tuna      378      25     3
5      Beef     1060      84    28
6    Frozen      339      15     4
7   Chicken      400      14     0
8   Chicken      286      25     3
9    Frozen      120      18     5
10   Frozen      260       5     3
```

Plotting

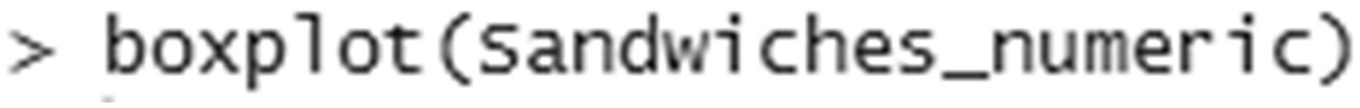
Graph a scatterplot between the variables



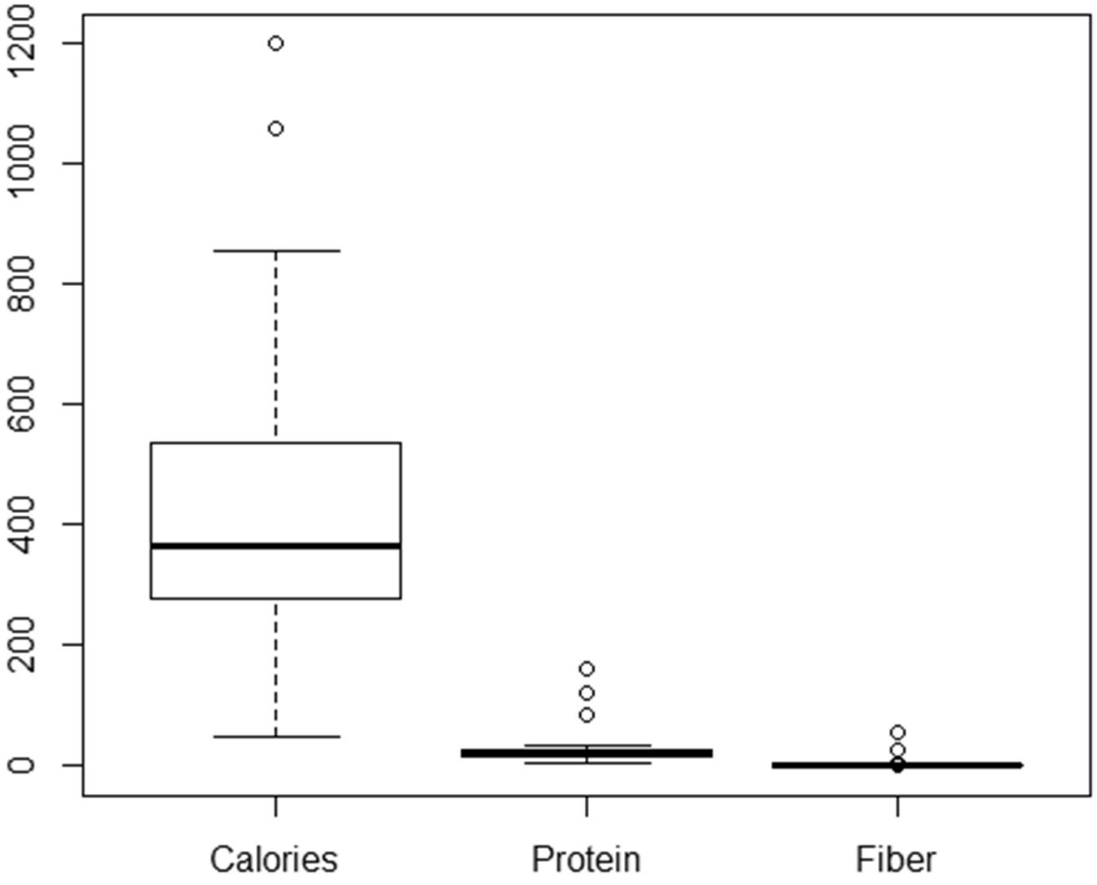
```
> Sandwiches_numeric <- select(Sandwiches, 2:4)
> pairs(Sandwiches_numeric)
```



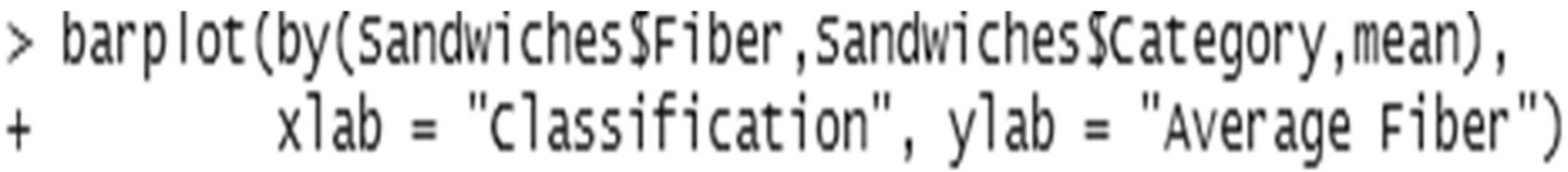
Identifying any outliers in the dataset



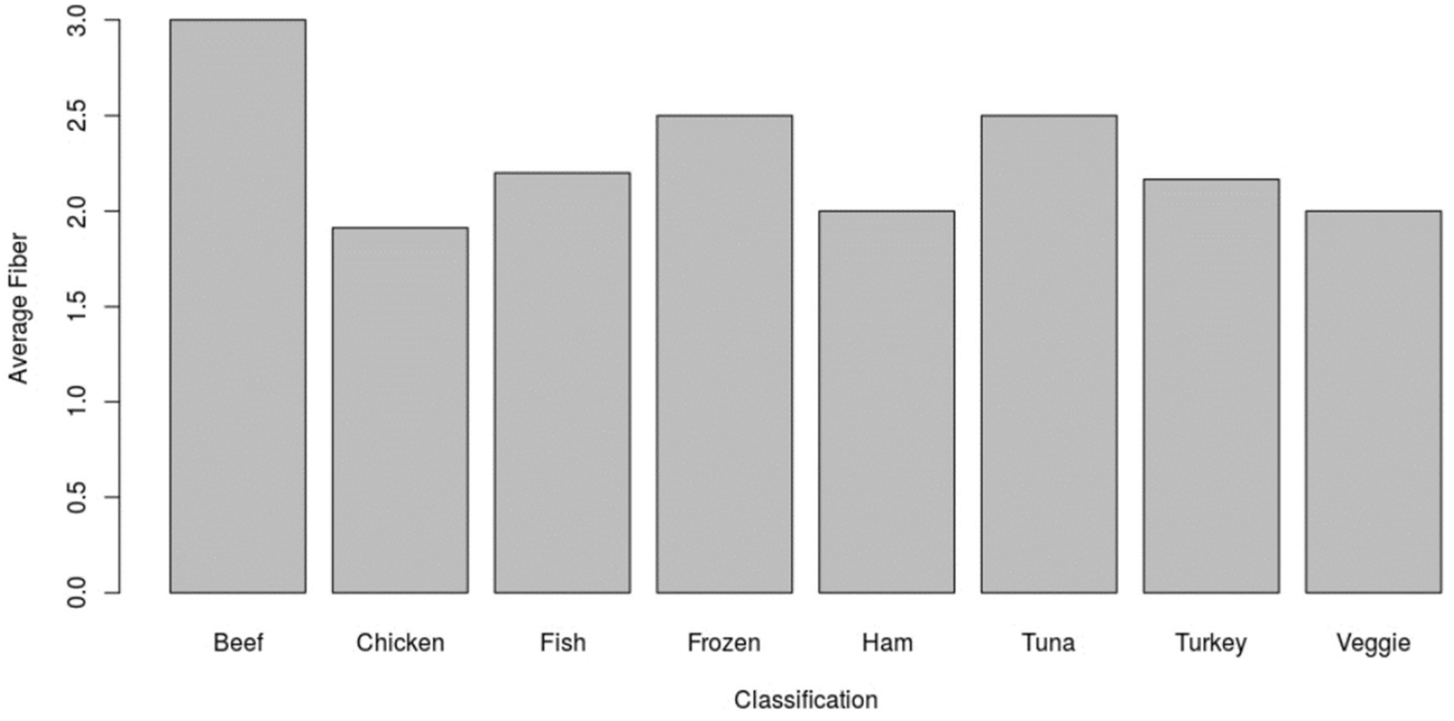
```
> boxplot(Sandwiches_numeric)
```



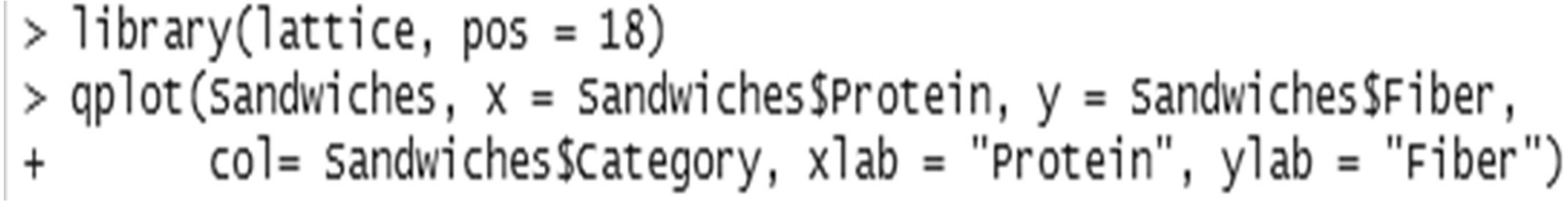
Create a bargraph of average value by Category



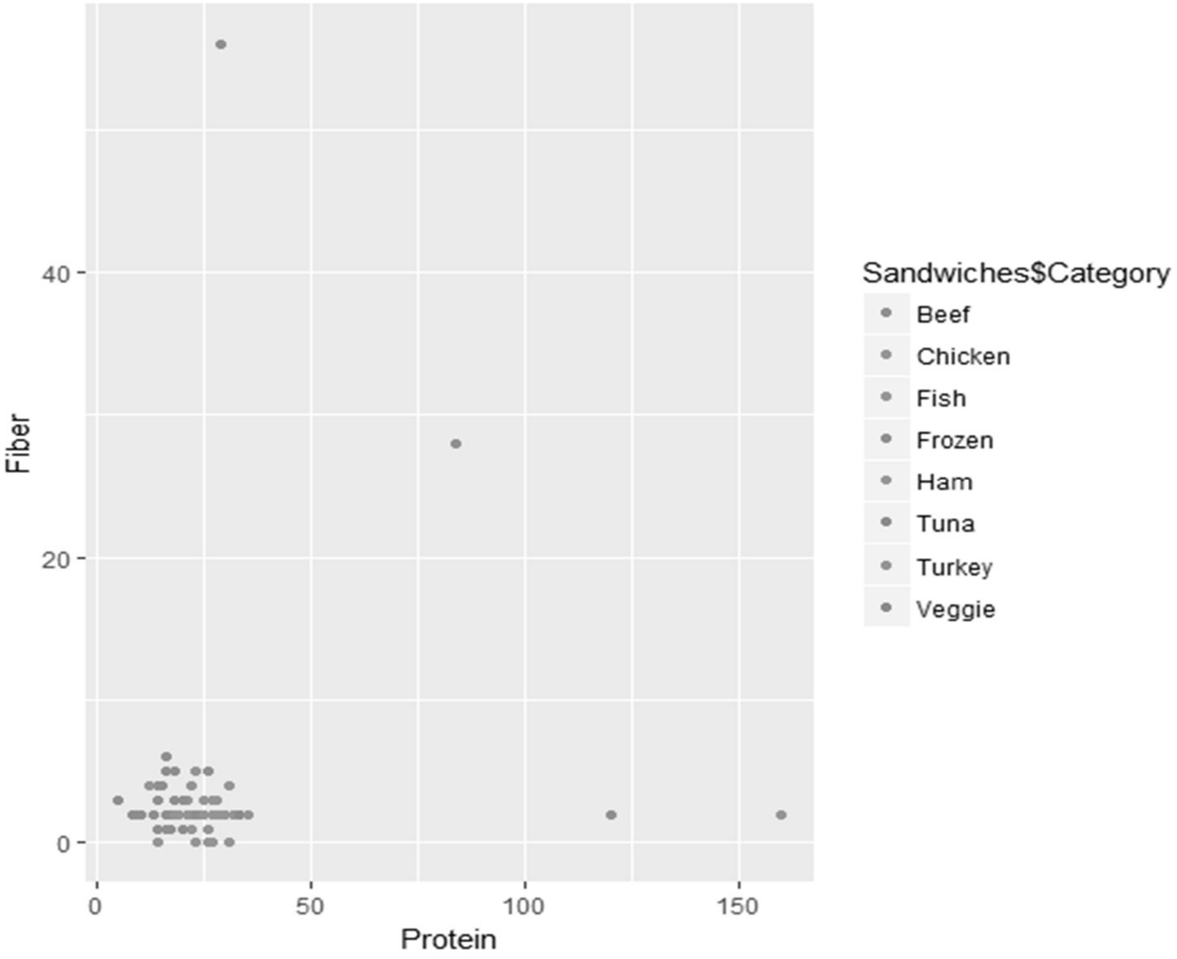
```
> barplot(by(Sandwiches$Fiber,Sandwiches$Category,mean),
+         xlab = "Classification", ylab = "Average Fiber")
```



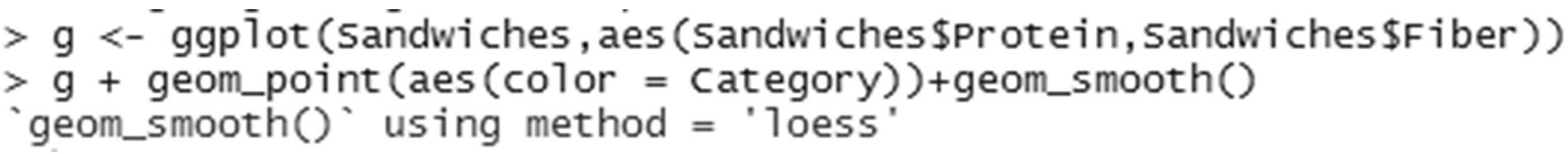
Create a scatterplot of Protein and Fiber, by Category



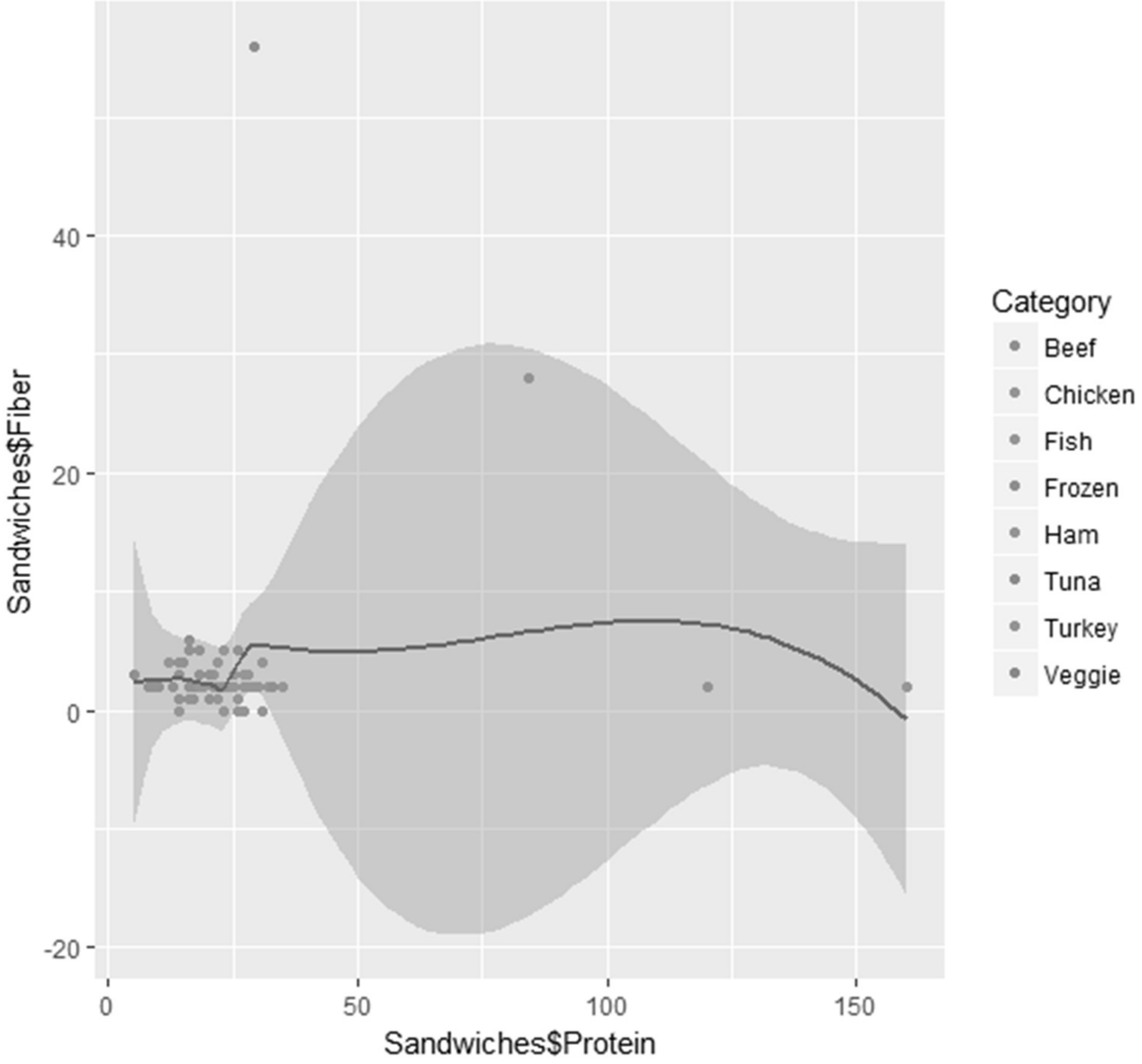
```
> library(lattice, pos = 18)
> qplot(Sandwiches, x = Sandwiches$Protein, y = Sandwiches$Fiber,
+       col= Sandwiches$Category, xlab = "Protein", ylab = "Fiber")
```



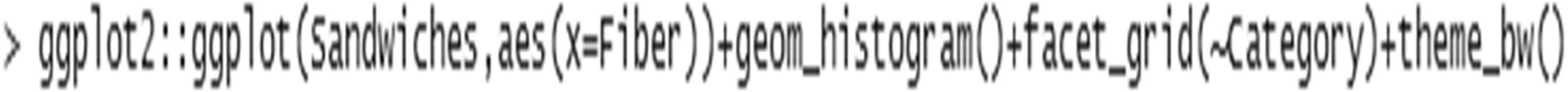
Add a smoothing line to the scatterplot



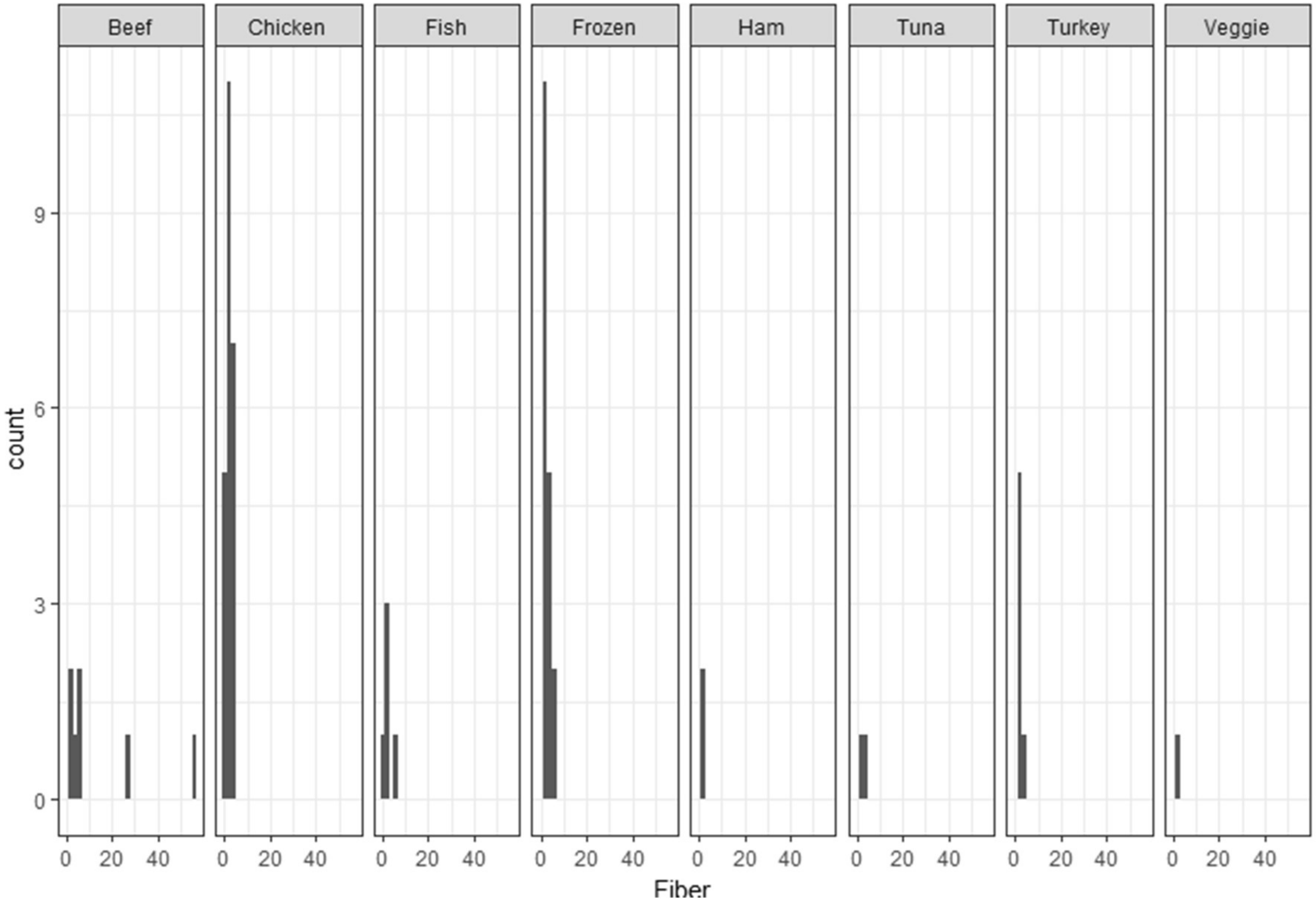
```
> g <- ggplot(Sandwiches,aes(Sandwiches$Protein,Sandwiches$Fiber))
> g + geom_point(aes(color = Category))+geom_smooth()
`geom_smooth()` using method = 'loess'
```



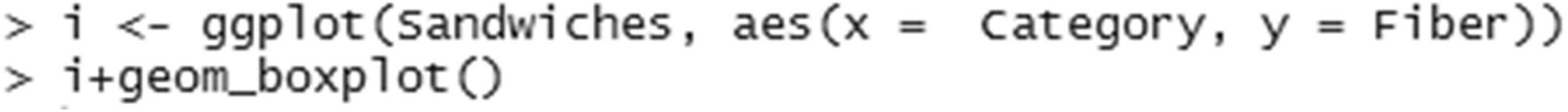
Create a histogram of Fiber, by Category



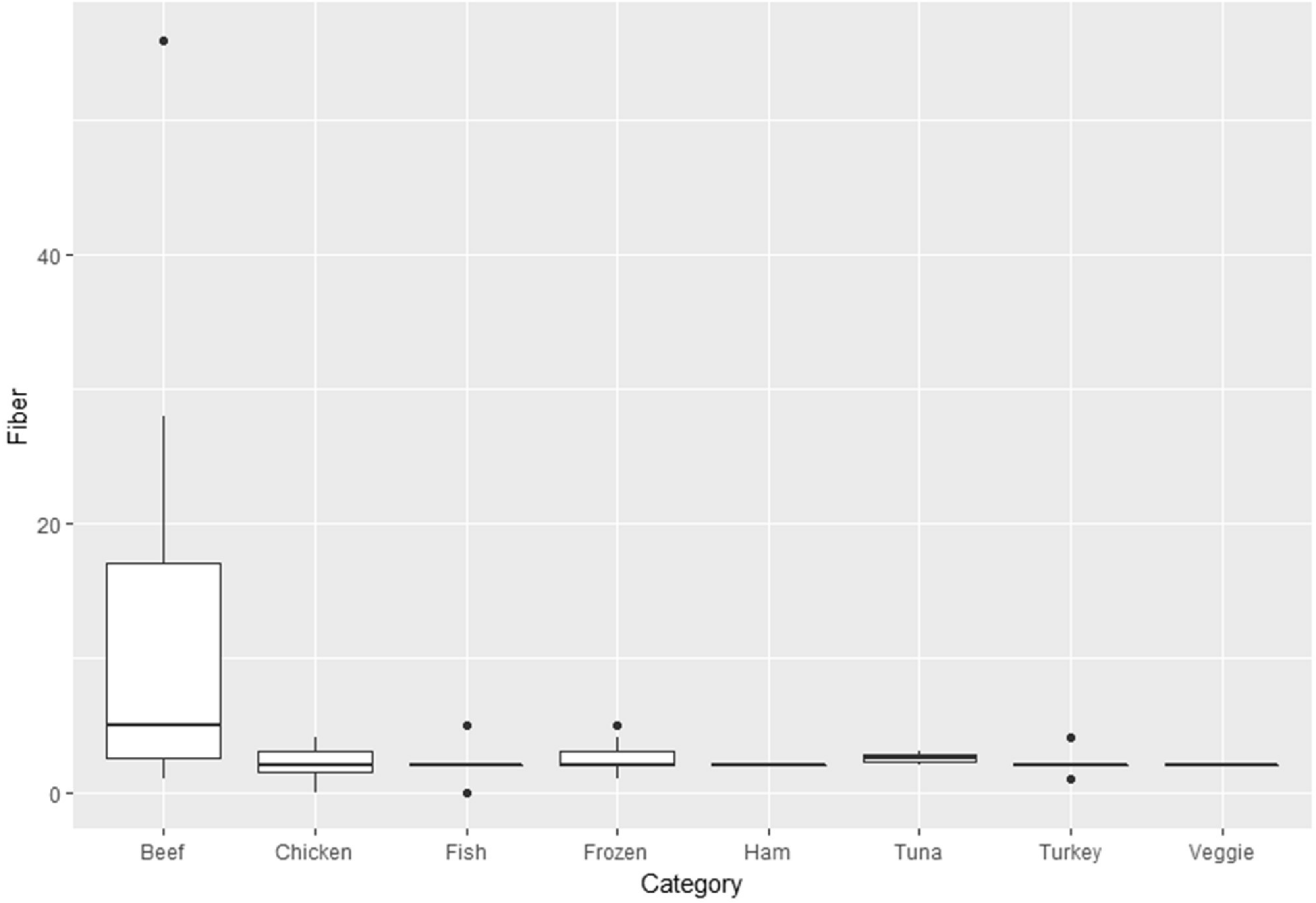
```
> ggplot2::ggplot(Sandwiches,aes(x=Fiber))+geom_histogram()+facet_grid(~Category)+theme_bw()
```



Create a boxplots of Fiber by Category



```
> i <- ggplot(Sandwiches, aes(x =  Category, y = Fiber))
> i+geom_boxplot()
```



Frozen?