# Week 2 Homework Steven Simonsen

Steven Simonsen

2024-01-26

### Ch.3 Homework: In this chapter, you can use the weight data set and perform all the actions covered here: selecting variables, filtering observations and reshaping.

```
weight_data <- read.csv("~/School/DSE5001 Intro to Data Science and Stats/Week 2/weight.csv")
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
## The following objects are masked from 'package:stats':
##
##     filter, lag
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(tidyr)
library(ggplot2)

#Included slicing data to reduce size of pdf down from hundreds of pages
weight_data %>%
  select(subjectid, gender, height,
                     weight, age, race) %>%
  rename(subject_id=subjectid) %>%
  filter(gender=='Female') %>%
  slice(1:10) %>%
  mutate(new_age = ifelse(age<30,'<30','>=30')) %>%
  arrange(desc(height))
```

```
##    subject_id gender height weight age race new_age
## 1       10042 Female  171.1   66.3  23    1     <30
## 2       10053 Female  170.7   83.7  44    2    >=30
## 3       10070 Female  167.1   76.0  23    6     <30
## 4       10038 Female  166.5   53.4  21    3     <30
## 5       10043 Female  166.0   78.2  22    2     <30
## 6       10061 Female  164.4   73.2  21    1     <30
## 7       10080 Female  159.0   68.4  37    1    >=30
## 8       10051 Female  157.2   88.6  45    1    >=30
## 9       10037 Female  156.0   65.7  26    2     <30
## 10      10077 Female  152.1   54.5  24    2     <30
```

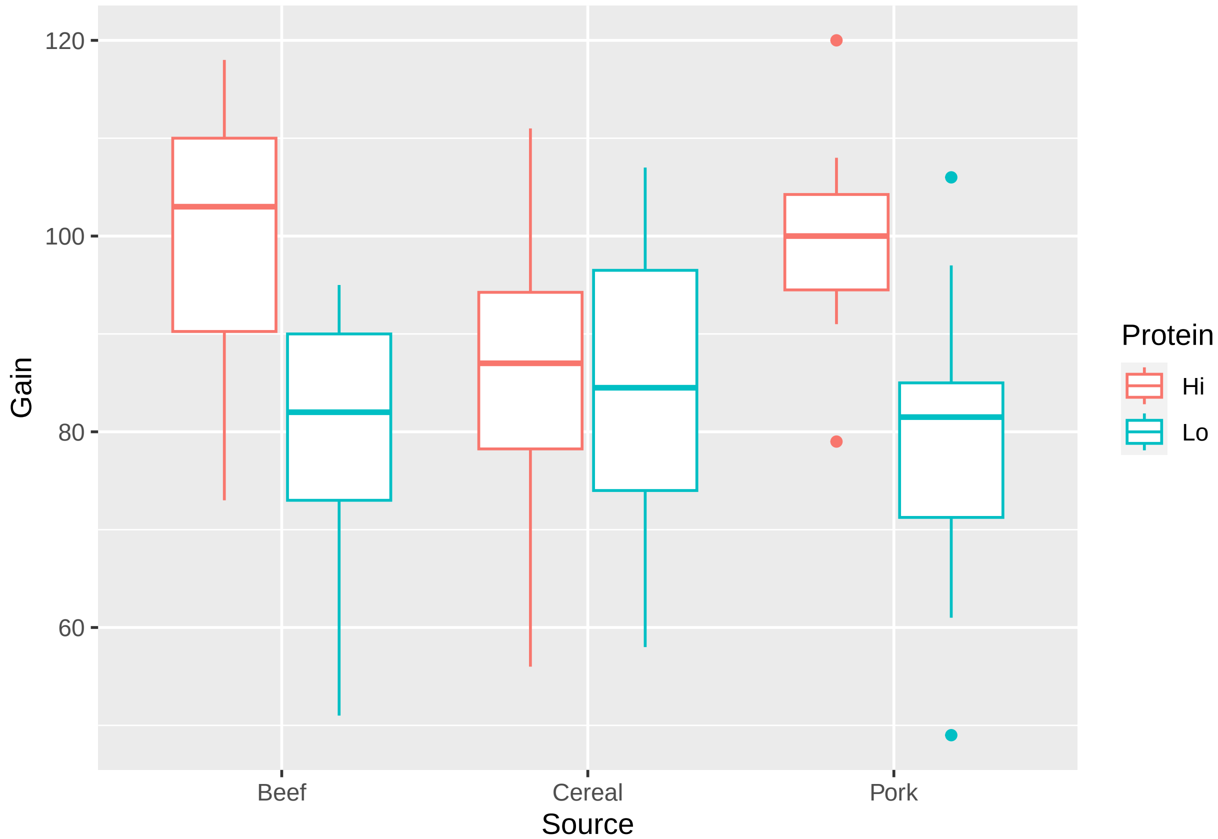
```
#Group By and Summarize
weight_data %>%
```

```r
  group_by(gender, race) %>%
  summarise(median_weight = median(weight))
```

```
## `summarise()` has grouped output by 'gender'. You can override using the
## `.groups` argument.

## # A tibble: 13 x 3
## # Groups:   gender [2]
##    gender  race median_weight
##    <chr>  <int>         <dbl>
##  1 Female     1          67
##  2 Female     2          68.6
##  3 Female     3          64
##  4 Female     4          60.3
##  5 Female     5          71.8
##  6 Female     6          59.6
##  7 Male       1          84.9
##  8 Male       2          86.6
##  9 Male       3          82.6
## 10 Male       4          73.7
## 11 Male       5          89.9
## 12 Male       6          75.9
## 13 Male       8          70
```

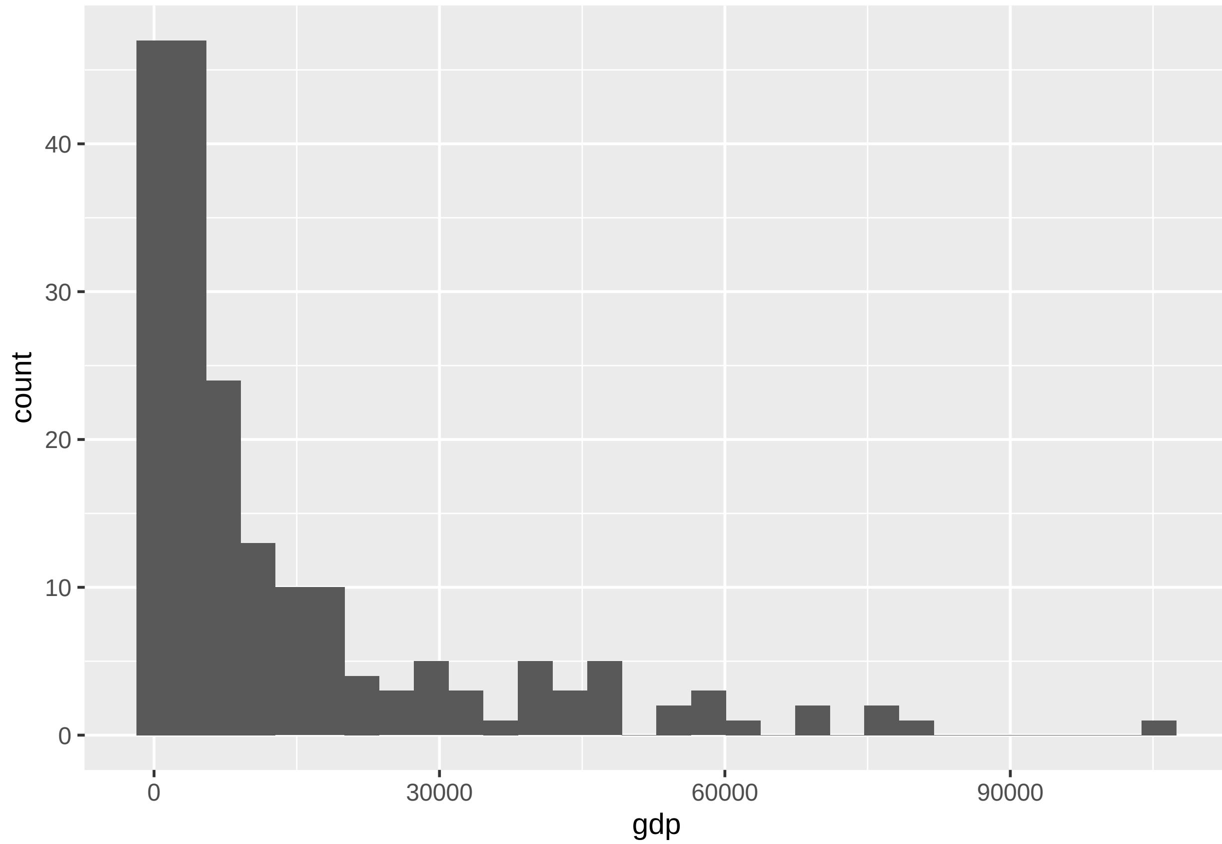
**Ch.4 Homework: As with the previous chapter on data wrangling, a valuable exercise based on this chapter is for the reader to use their own data-sets to practice with all the plotting methods that are described in the chapter. It may be that different data sets may be required for different types of plots. See additional datasets below**

```r
FatRats <- read.csv("~/School/DSE5001 Intro to Data Science and Stats/Week 2/FatRats.csv")
nominal_gdp_per_capita <- read.csv("~/School/DSE5001 Intro to Data Science and Stats/Week 2/nominal_gdp
quartet <- read.csv("~/School/DSE5001 Intro to Data Science and Stats/Week 2/quartet.csv")
sleepstudy <- read.csv("~/School/DSE5001 Intro to Data Science and Stats/Week 2/sleepstudy.csv")
TitanicSurvival <- read.csv("~/School/DSE5001 Intro to Data Science and Stats/Week 2/TitanicSurvival.cs

#FatRats data vis
FatRats %>%
  ggplot() +
  geom_boxplot(aes(x=Source, y=Gain, color=Protein))
```

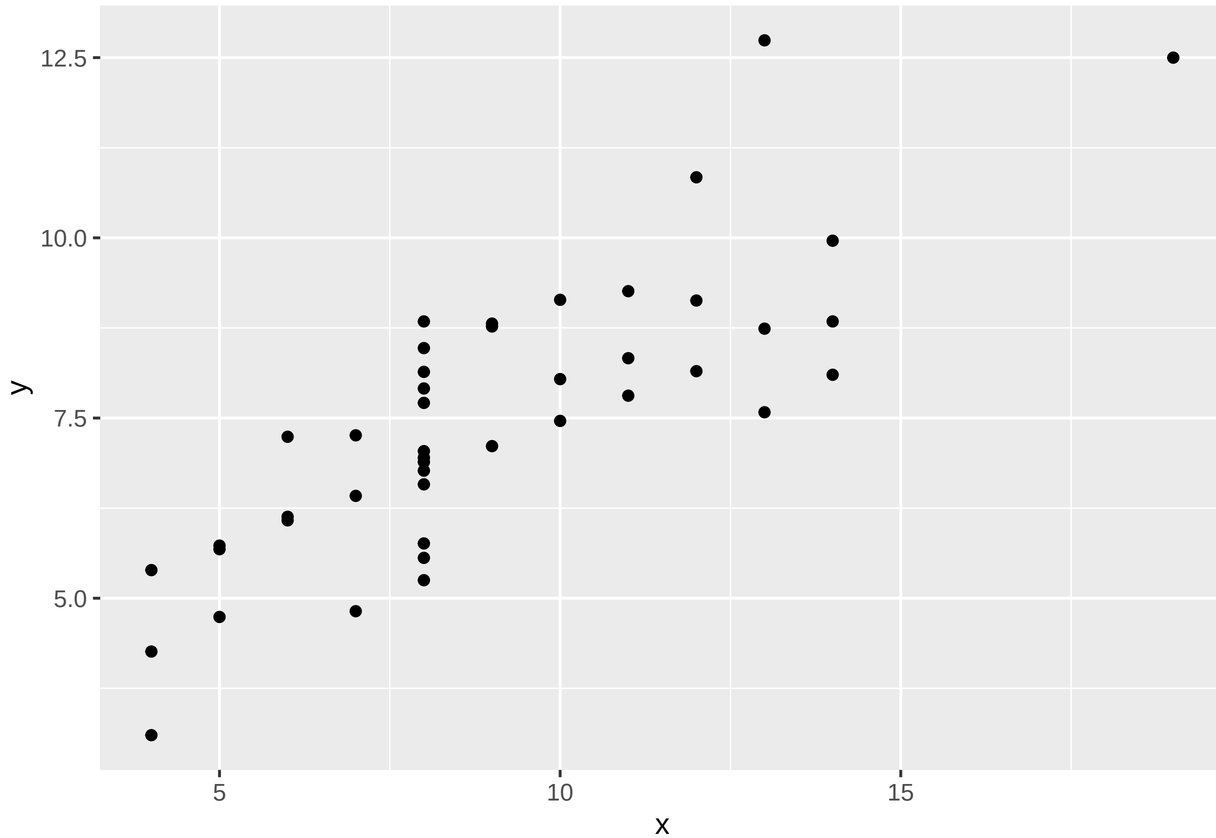



```r
#nominal_gdp_per_capita vis
nominal_gdp_per_capita %>%
  ggplot() +
  geom_histogram(aes(x=gdp))
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

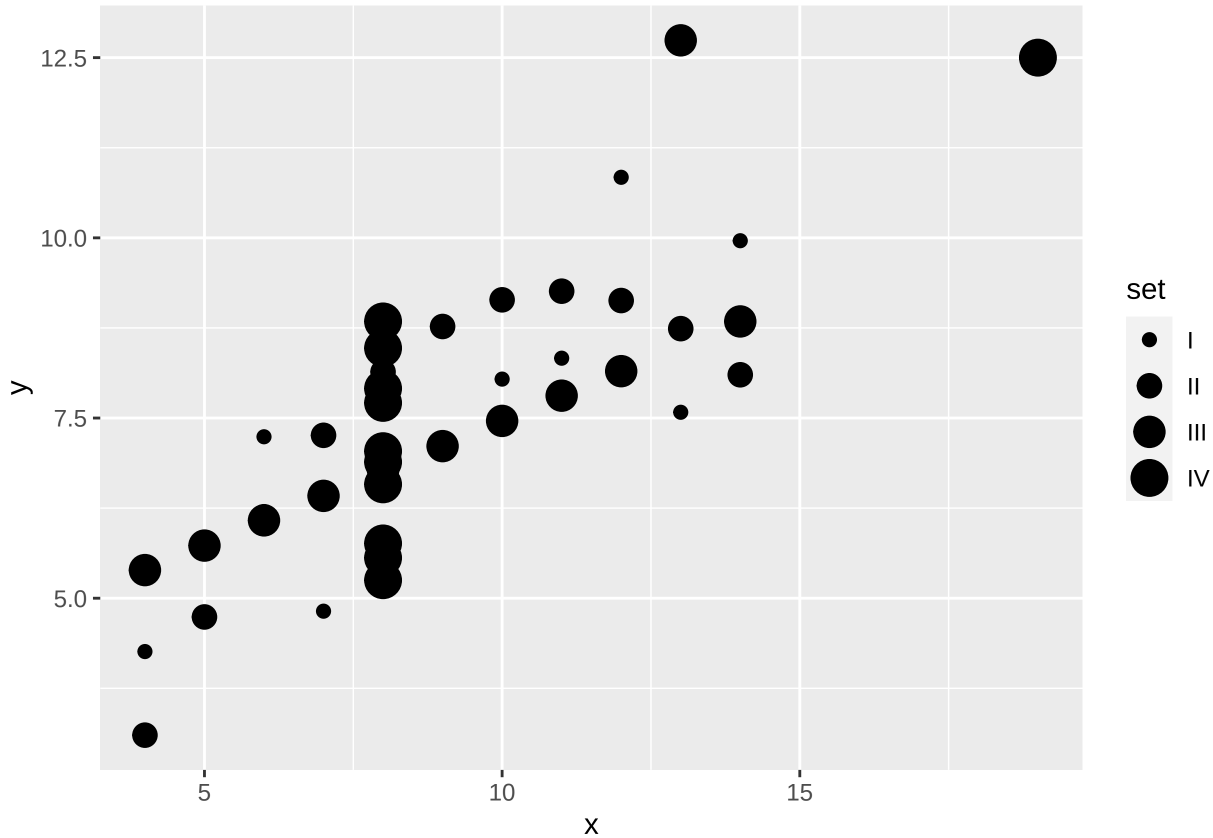



```
#quartet
quartet %>%
  ggplot() +
  geom_point(aes(x=x, y=y))
```



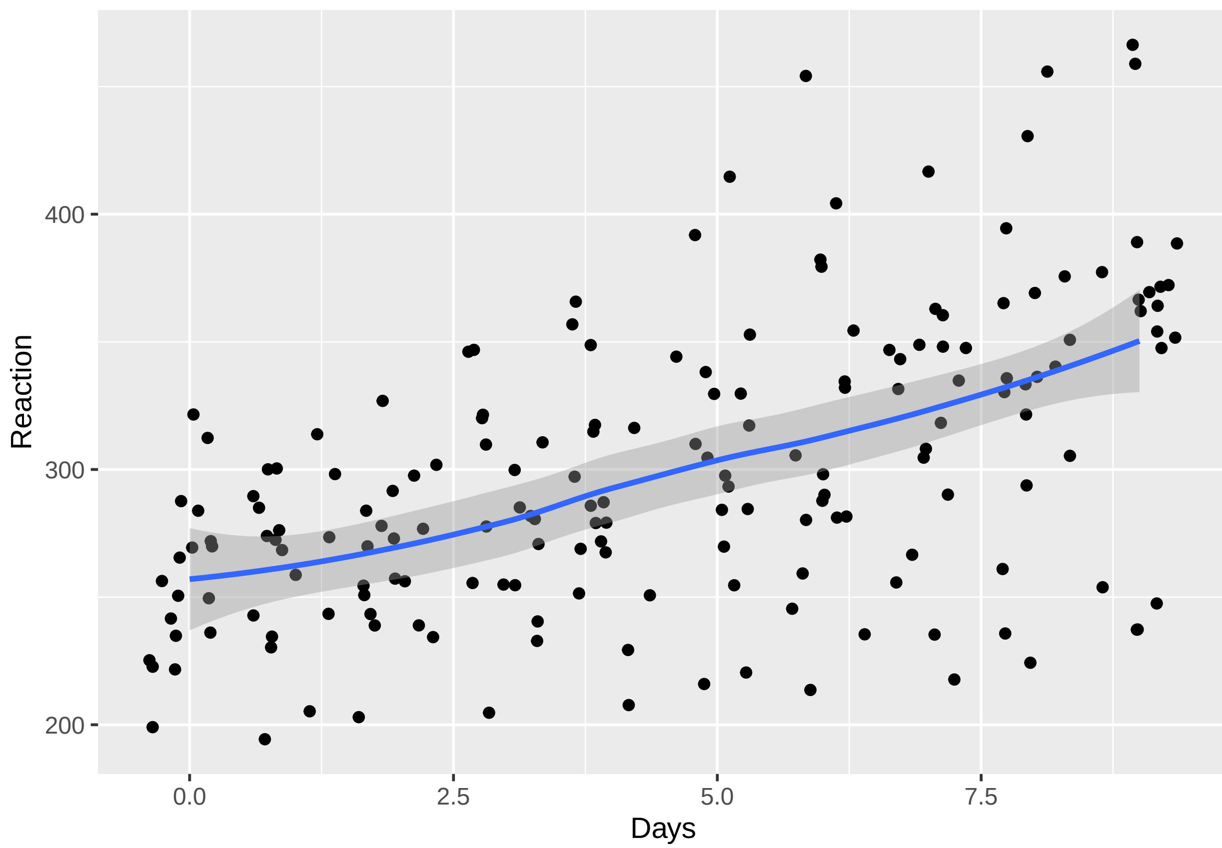

```
#Set size by set
quartet %>%
  ggplot() +
  geom_point(aes(x=x, y=y, size=set))
```

```
## Warning: Using size for a discrete variable is not advised.
```



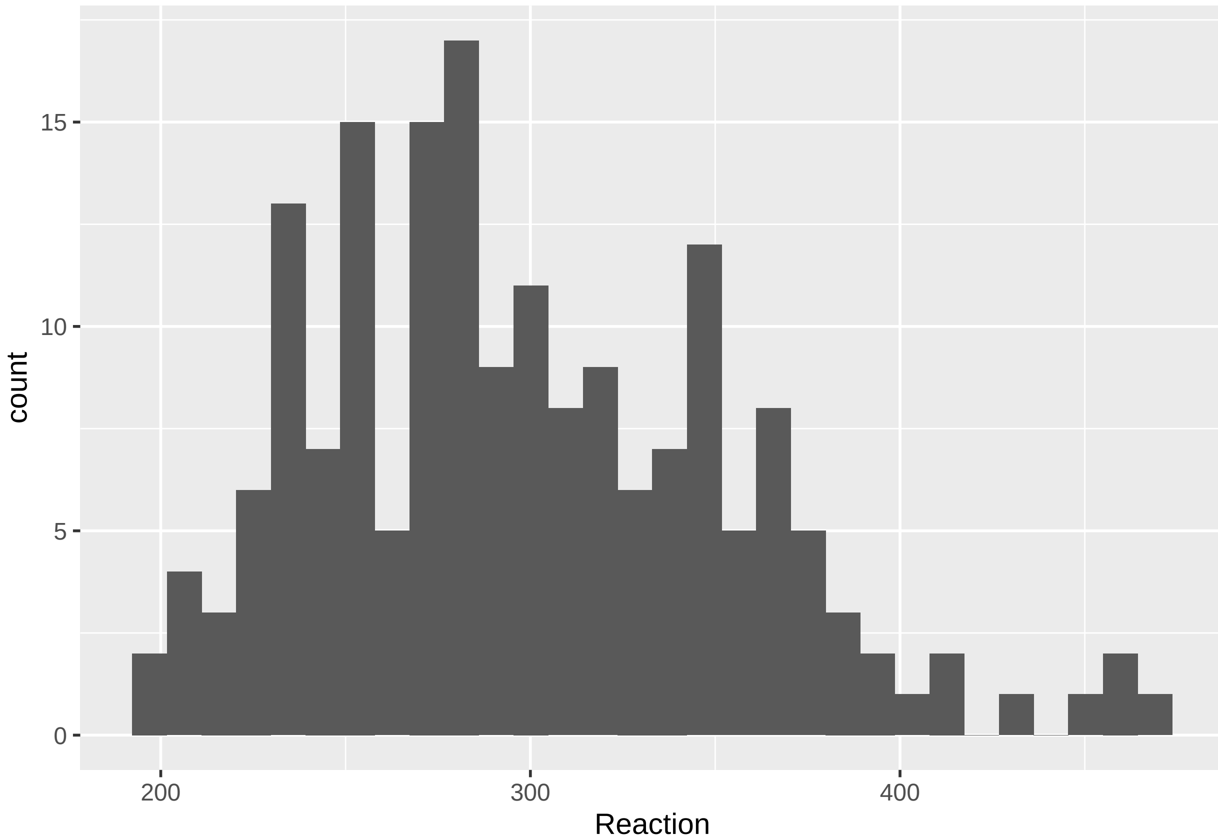

```
#Sleepstudy
#geom_smooth showed a fairly linear behavior correlated between days and reaction
sleepstudy %>%
  ggplot() +
  geom_jitter(aes(x=Days, y=Reaction)) +
  geom_smooth(aes(x=Days, y=Reaction))
```

```
## `geom_smooth()` using method = 'loess' and formula = 'y ~ x'
```

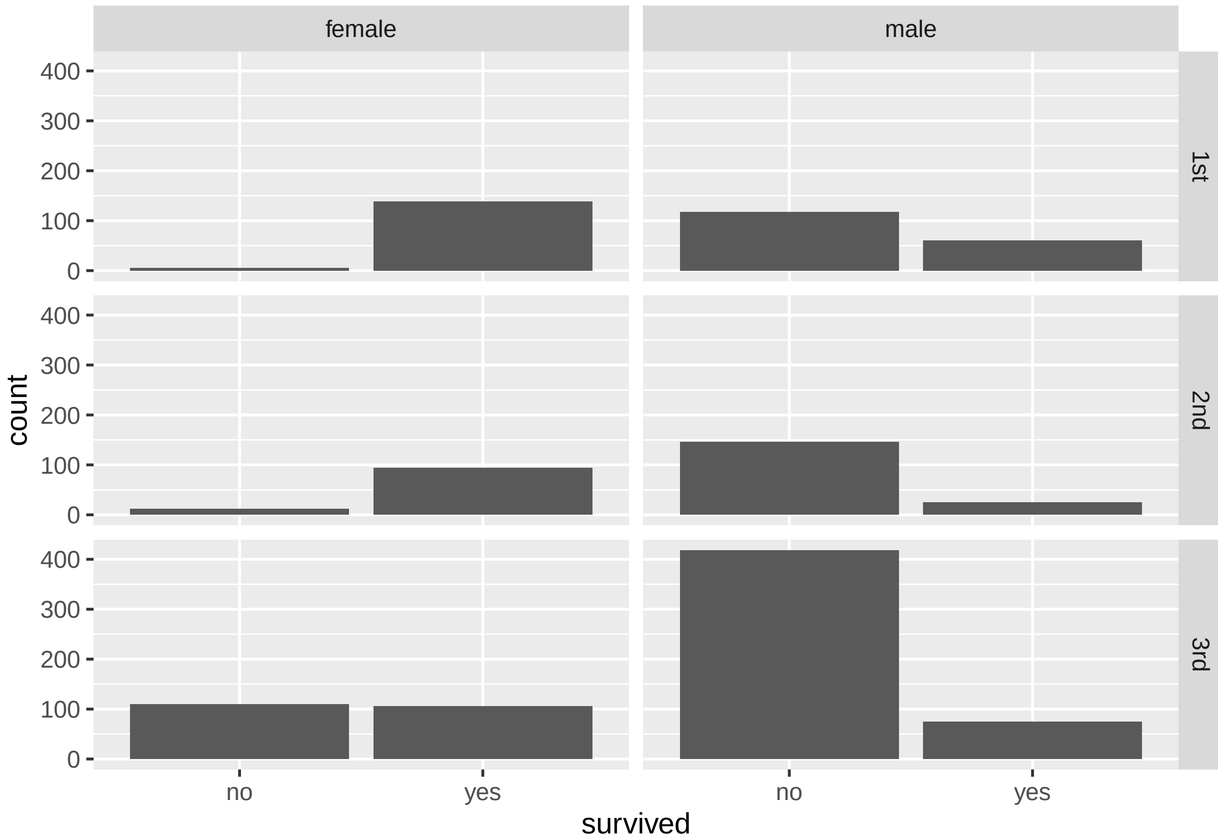



```r
#Histogram shows that most people tend to react the most between 225-300
sleepstudy %>%
  ggplot() +
  geom_histogram(aes(x=Reaction))
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```




```r
#TitanicSurvival

#Use of bar and facet plot to break down survival.
TitanicSurvival %>%
  ggplot()+
  geom_bar(aes(x=survived))+
  facet_grid(cols = vars(sex), rows = vars(passengerClass))
```



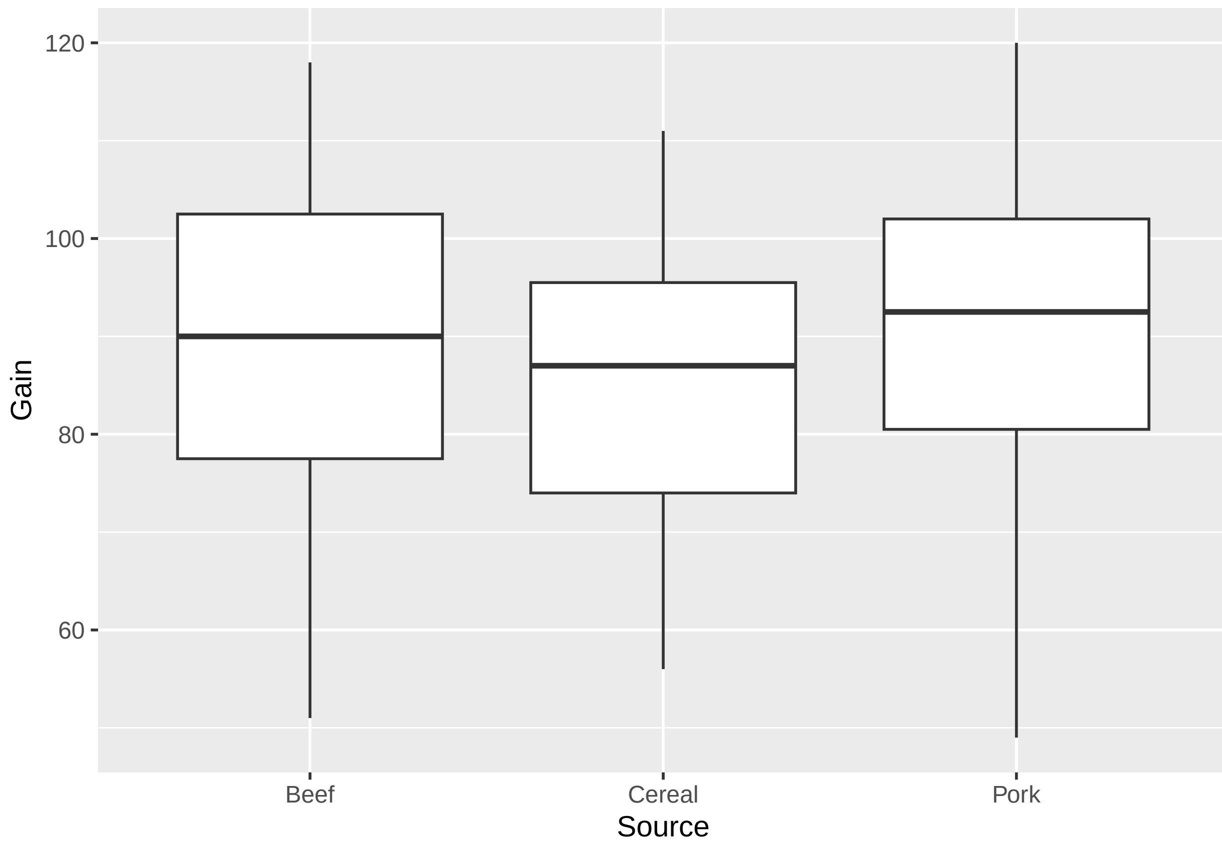

**Ch. 5 Homework: Perform some univariate exploratory analyses. For example, from one or more variables of interest plot their histograms and boxplots, both overall, and when the variable is grouped according to values of another variable. In parallel, calculate summary statistics measures of central tendency, dispersion, skewness, and kurtosis. Compare the plots to the tables of quantities to be able to get a grasp on how certain summaries of the data manifest themselves visualization, and how certain properties of the plots manifest themselves in summary statistics. For example, see whether histograms with long tails correspond to relatively high values of skewness, and vice versa.**

```
#Using FatRats Dataset
#Univariate Exploration


#univariate exploration by source
gbs <- FatRats %>%
  group_by(Source)

gbs %>%
  summarise(mean_Gain = mean(Gain),variance_Gain = var(Gain),
            stddeviation_Gain = sd(Gain),
            median_Gain = median(Gain),
            maximum_value = max(Gain),
            minimum_value = min(Gain),
```

```r
            InterQuartileRange = IQR(Gain))
```

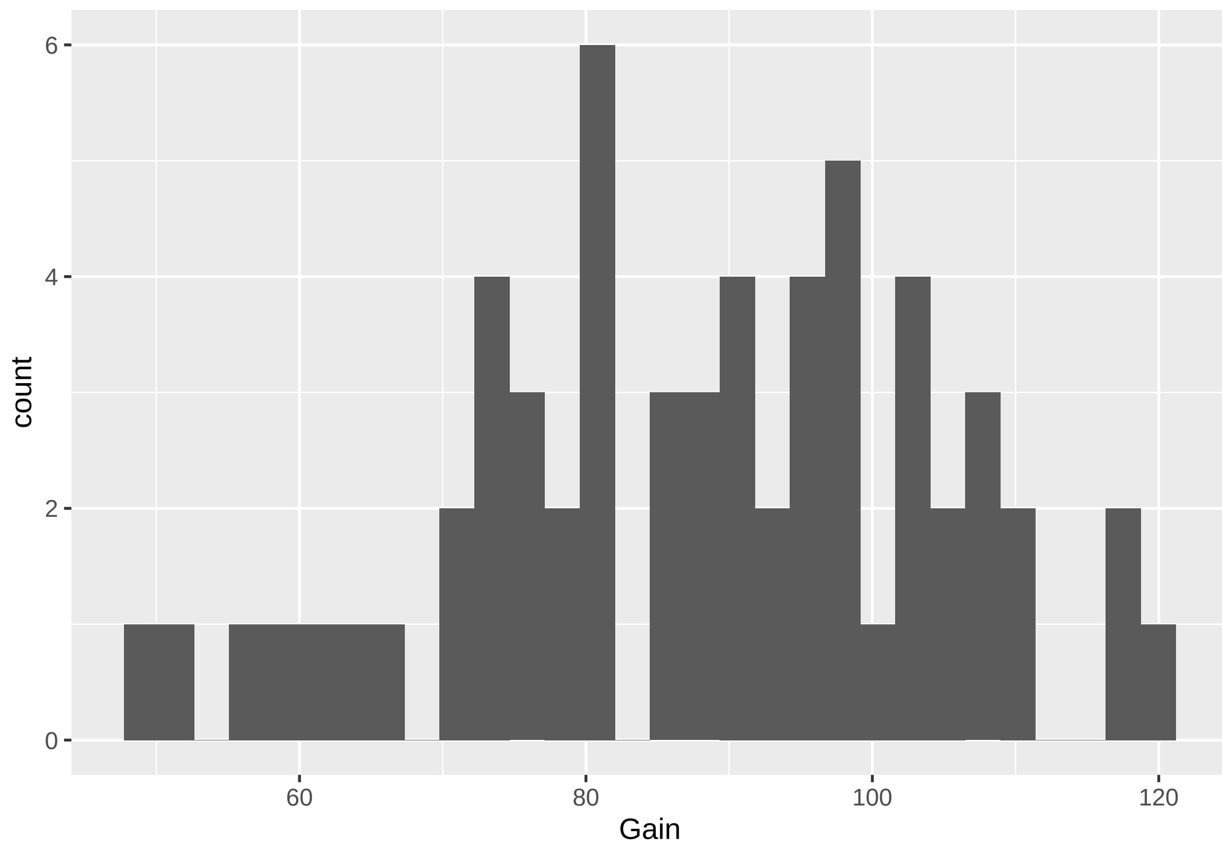
```
## # A tibble: 3 x 8
##   Source mean_Gain variance_Gain stddeviation_Gain median_Gain maximum_value
##   <chr>      <dbl>         <dbl>             <dbl>       <dbl>         <int>
## 1 Beef        89.6          314.              17.7        90             118
## 2 Cereal      84.9          225.              15.0        87             111
## 3 Pork        89.2          297.              17.2        92.5           120
## # i 2 more variables: minimum_value <int>, InterQuartileRange <dbl>
```

```r
#Boxplot showing source vs. Gain
FatRats %>%
  ggplot() +
  geom_boxplot(aes(x=Source, y=Gain))
```




```r
#Histogram showing count by Gain
FatRats %>%
  ggplot()+
  geom_histogram(aes(x=Gain))
```

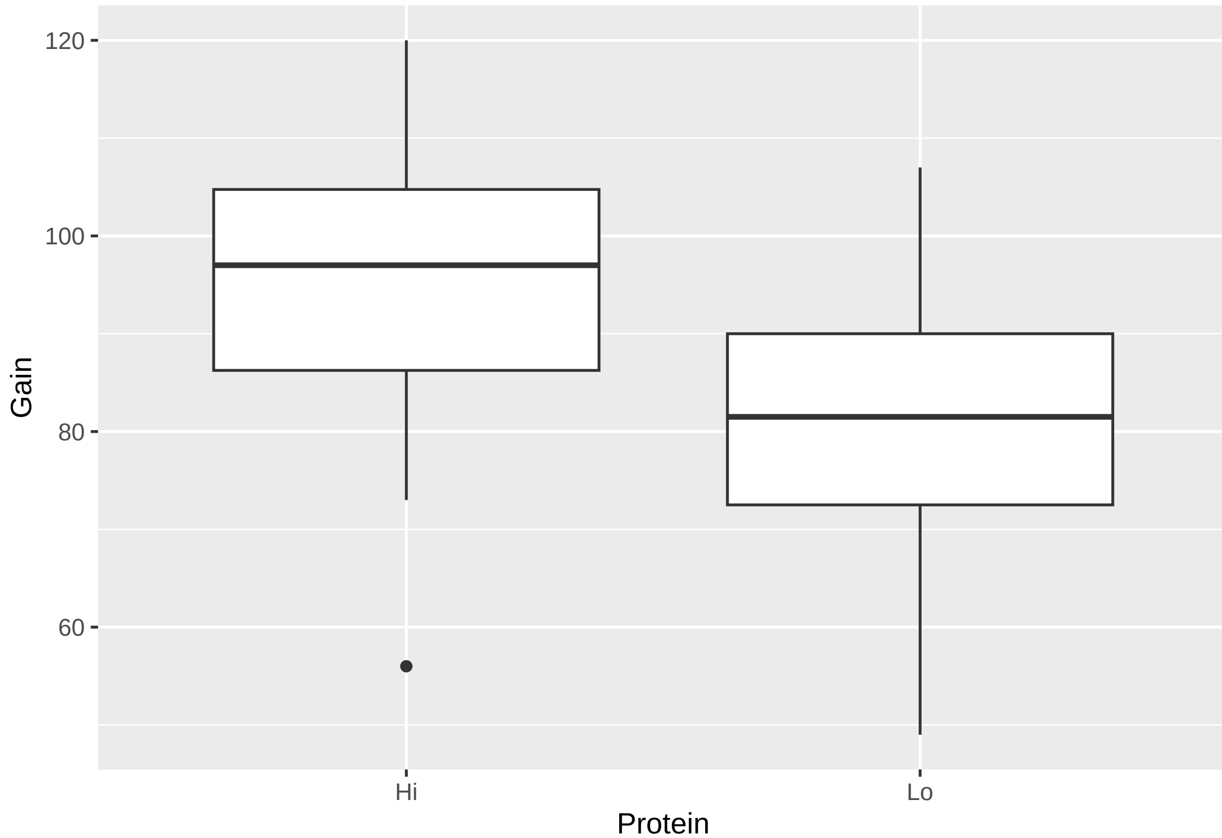
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```




```r
#Group Data by Protein
gbp <- FatRats %>%
  group_by(Protein)

gbp %>%
  summarise(mean_Gain = mean(Gain),variance_Gain = var(Gain),
            stddeviation_Gain = sd(Gain),
            median_Gain = median(Gain),
            maximum_value = max(Gain),
            minimum_value = min(Gain),
            InterQuartileRange = IQR(Gain))
```

```
## # A tibble: 2 x 8
##   Protein mean_Gain variance_Gain stddeviation_Gain median_Gain maximum_value
##   <chr>       <dbl>         <dbl>             <dbl>       <dbl>         <int>
## 1 Hi           95.1          222.              14.9        97             120
## 2 Lo           80.7          226.              15.0        81.5           107
## # i 2 more variables: minimum_value <int>, InterQuartileRange <dbl>
```

```r
FatRats %>%
  ggplot() +
  geom_boxplot(aes(x=Protein, y=Gain))
```




```r
#Measures of Central Tendency
trimmed_mean <- function(x,trim=0.1){
  n <- length(x)
  lo <- floor(n*trim)+1
  hi <- n+1-lo
  sort(x)[lo:hi] %>%
    mean()
}
print(paste("The trimmed mean of the Gain of the Fatrats dataset is: ",trimmed_mean(FatRats$Gain)))
```

```
## [1] "The trimmed mean of the Gain of the Fatrats dataset is:  88.5416666666667"
```

```r
iqr_mean <- function(x){
  q1 <- quantile(x,probs=0.25)
  q2 <- quantile(x,probs=0.75)
  x[x>q1 & x < q2] %>%
    mean()
}

print(paste("The IQR Mean of the Gain of the Fatrats dataset is: ",iqr_mean(FatRats$Gain)))
```

```
## [1] "The IQR Mean of the Gain of the Fatrats dataset is:  88.6666666666667"
```

```r
winsorized_mean <- function(x,trim=0.1){
  low <- quantile(x,probs=trim)
  high <- quantile(x,probs=1-trim)
  x[x<low]<- low
  x[x>high]<-high
```

```r
  mean(x)
}

print(paste("The winsorized_mean of the Gain of the Fatrats dataset is: ",winsorized_mean(FatRats$Gain)))
```

```
## [1] "The winsorized_mean of the Gain of the Fatrats dataset is:  88.2133333333333"
```

```r
#Skewness
skewness <- function(x,dof=1){
  xbar <- mean(x)
  s <- sd(x)
  mean((x-xbar)^3)/s^3
}

print(paste("The skewness of the Gain of the Fatrats dataset is: ",skewness(FatRats$Gain)))
```

```
## [1] "The skewness of the Gain of the Fatrats dataset is:  -0.282497163714679"
```

```r
#Kurtosis
kurtosis <- function(x){
  z <- (x-mean(x))/sd(x)
  mean(z^4)
}

print(paste("The kurtosis of the Gain of the Fatrats dataset is: ",kurtosis(FatRats$Gain)))
```

```
## [1] "The kurtosis of the Gain of the Fatrats dataset is:  2.55883965202515"
```

##Explanation of Ch.5: Findings: I examined the FatRats dataset, which measures the gain of what I assume to be a rat's weight over a given time period by measuring against two variables. The two variables within the study include a high/low protein intake, and different sources of food (ex: beef). I first performed univariate discovery by grouping the dataset by the Source variable. Overall, beef produced the highest mean gain. However, pork produced the higher median gain, along with the highest maximum value. This leads me to believe that although pork may have had some higher outliers, beef produced the highest consistent Gain since the mean was the highest. I also performed univariate discovery by grouping the dataset by the high/low variable. Overall, a high protein intake produced a larger gain.

Additional data visualization was performed to affirm my conclusions. Sure enough, when visualizing the Source (x-axis) and the Gain (y-axis), the pork source had longer tails than the beef source. This supports the finding of a larger spread and therefore a higher median. An additional finding showed that cereal appeared to be the most negatively skewed in comparison to the Beef and Pork sources. When comparing the Protein (x-axis) variable in comparison to Gain (y-axis), there was an outlier on the high Protein variable. Additional findings related to this outlier may be present if this individual case were to be examined in further detail.

Finally, the measures of central tendency, skewness, and kurtosis were examined. Measurements such as the trimmed mean, the winsorized mean, and the interquartile mean of the Gain were all very similar measurements at around 88. Although I was not given units of measurement for this study, I assume this to be calculated in grams. In comparison to the mean relative to the various sources, this is consistent with beef and pork, and about 4 grams higher than the mean Gain relative to cereal. The skewness of the gain was calculated to be -0.282497163714679, meaning the data is close to symmetrical relative to Pearson's Second Coefficient. This also means the horizontal pull on the data is minimal. Kurtosis of the Gain was also measured and calculated to be a 2.55883965202515. This leaves the Excess Kurtosis with a value of 3-2.55883965202515=0.4411603. Overall, given that the kurtosis is close to 3, the vertical pull is minimal in the dataset, with a fairly normal peak in the data.