Some initial summarization and plotting of data can often reveal errors. In this example, we continue with some of the WDFW fisheries data for 2014.

library(tidyr)  
library(dplyr)

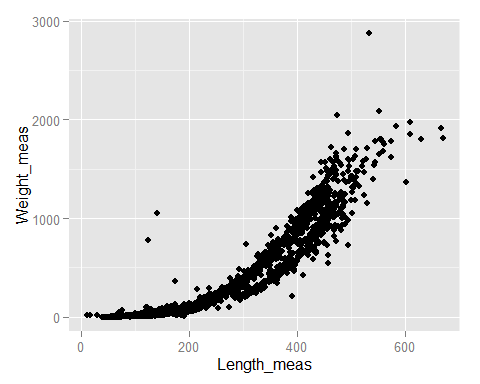
##   
## Attaching package: 'dplyr'  
##   
## The following object is masked from 'package:stats':  
##   
## filter  
##   
## The following objects are masked from 'package:base':  
##   
## intersect, setdiff, setequal, union

library(ggplot2)  
df <- read.csv("D:/fishdata.csv", header = TRUE, sep = ",")

Initially, we create an xy plot of Weight x Length for the entire dataset and including all species.

qplot(Length\_meas, Weight\_meas, data = df)

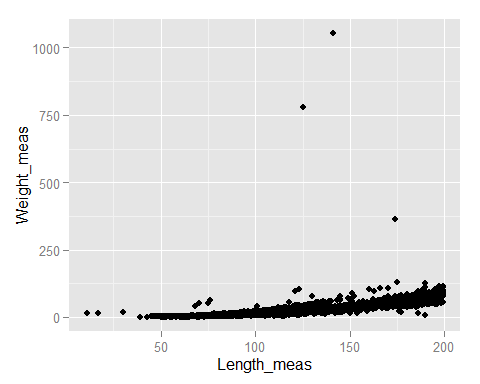
## Warning: Removed 338 rows containing missing values (geom\_point).



It's clear that there are some outliers at the lengths <200. So we'll filter the data and replot for these smaller lengths.

df200 <- df %>%  
 filter(Length\_meas < 200)  
qplot(Length\_meas, Weight\_meas, data = df200)

## Warning: Removed 335 rows containing missing values (geom\_point).



This chart makes it clear that there are 3 unusual observations. Often these are due to clerical errors (e.g., decimal in the wrong place).

We can again filter the dataset to extract the outlying observations.

dfoutliers <- df %>%  
 filter(Length\_meas < 200) %>%  
 filter(Weight\_meas > 250)  
dfoutliers

## WWSurveyId\_id SurveyNum\_id FishSampNum\_id Species\_code Length\_meas  
## 1 10167 113 1 LW 125  
## 2 10167 117 9 TNC 141  
## 3 10167 120 10 BBH 174  
## FishLenMo\_code Weight\_meas  
## 1 NA 780  
## 2 NA 1054  
## 3 NA 363

The outlying points can either be corrected or removed from the dataset as appropriate.There may be additional outliers in the data for measurements >200, but this won't be clear until the data are separated by species.

library(lattice)  
species.f <- factor(df$Species\_code)  
xyplot(df$Weight\_meas~df$Length\_meas|species.f)

