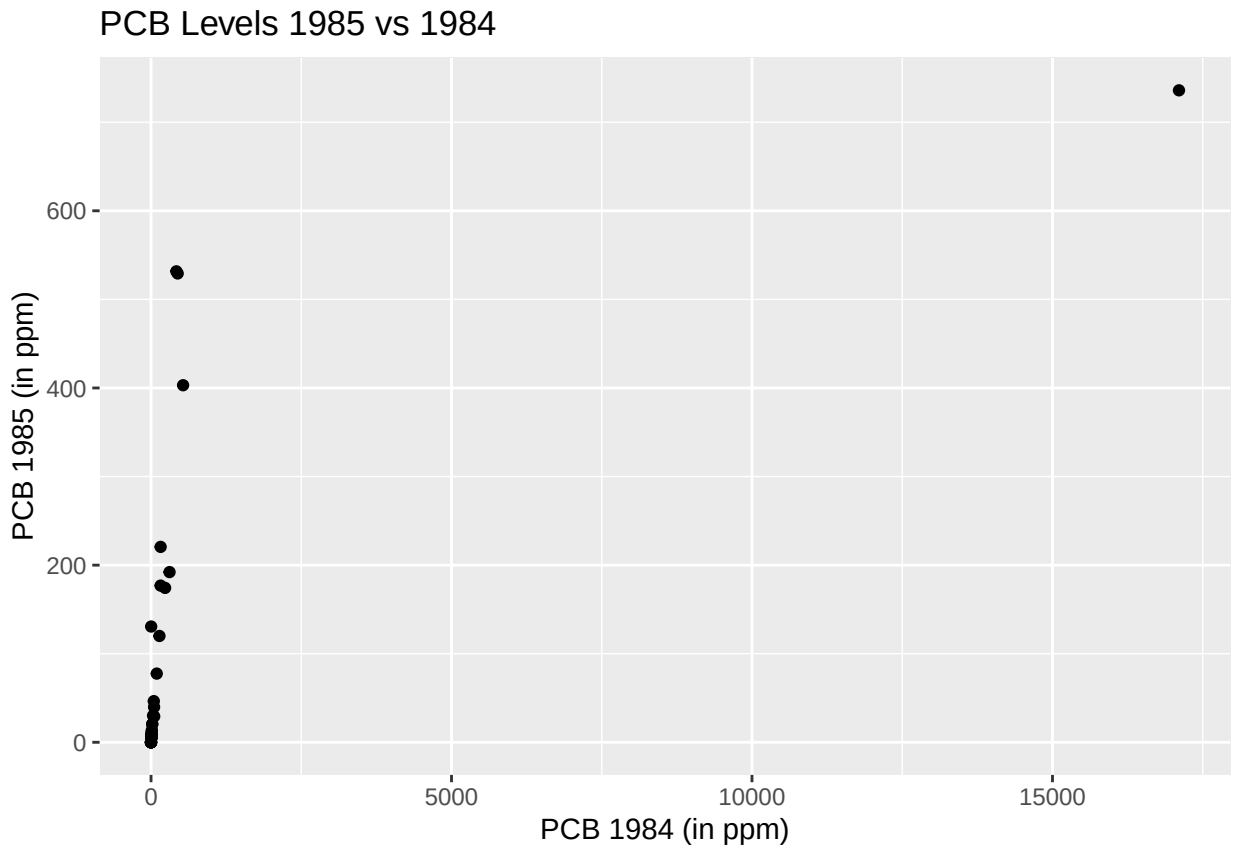
Stat230 - HW 3

Owen Forman

```
##
## 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
```

2a) **Answer:** The extreme outlier case was Boston Harbor, which had extremely high pcb levels in 1984. Upon removing the outlier though, it still turns out that the data is not yet suited for a SLR model. We can see this visually in the scatter plot "PCB Levels 1985 vs 1984 (No Outlier)" since the trend does not appear to be linear. Additionally, after attempting to fit a SLR model, it becomes even clearer that the model is not appropriate, as the residual plot"Pcb Resid plot (No outliers)" shows clear grouping around x = 0 and residuals become more widespread at larger x's which is a clear patter. Thus an SLR model cannot be used despite removing the outlier.

```
#reads in data
pcb <- read.csv("https://www.math.carleton.edu/ckelling/data/Pcb.csv")

#creates scatterplot of pcb85 vs pcb84.
ggplot(data = pcb, aes(x = pcb84, y = pcb85)) +
  geom_point() +
  labs(title = "PCB Levels 1985 vs 1984", x = "PCB 1984 (in ppm)", y ="PCB 1985 (in ppm)")
```



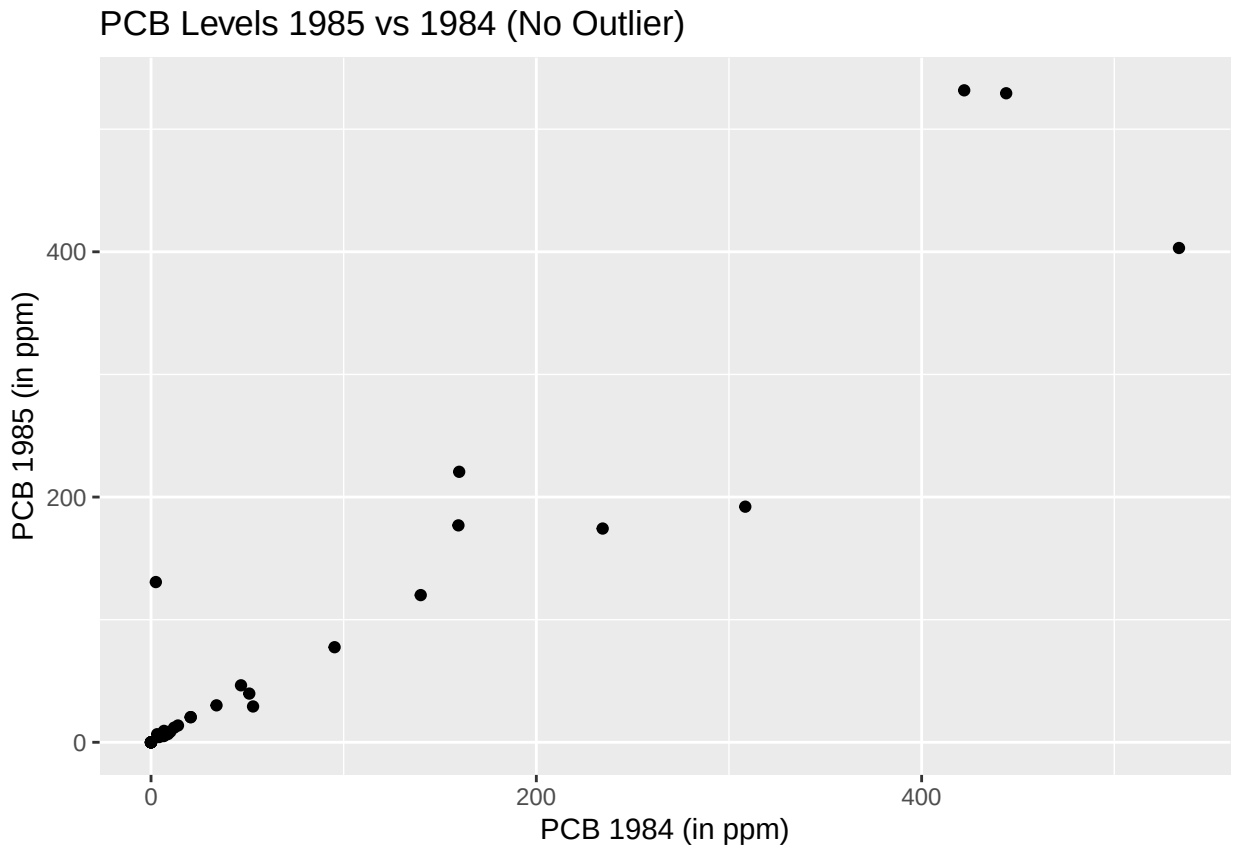

```
#identifies extreme outlier value
which(pcb$pcb84 > 10000)
```

```
## [1] 4
```

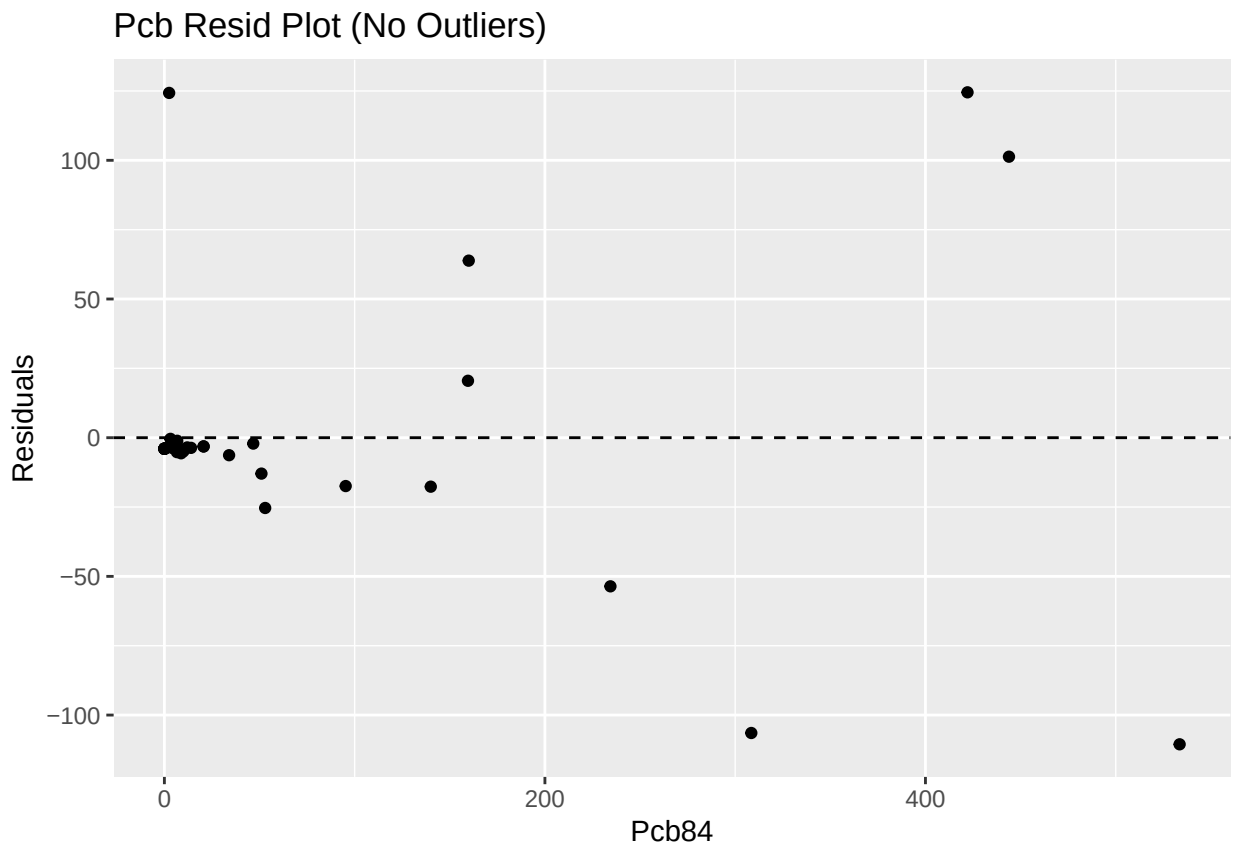
```
#prints row with outlier value
pcb[4,]
```

```
##            Site    pcb84 pcb85
## 4 Boston Harbor 17104.86   736
```

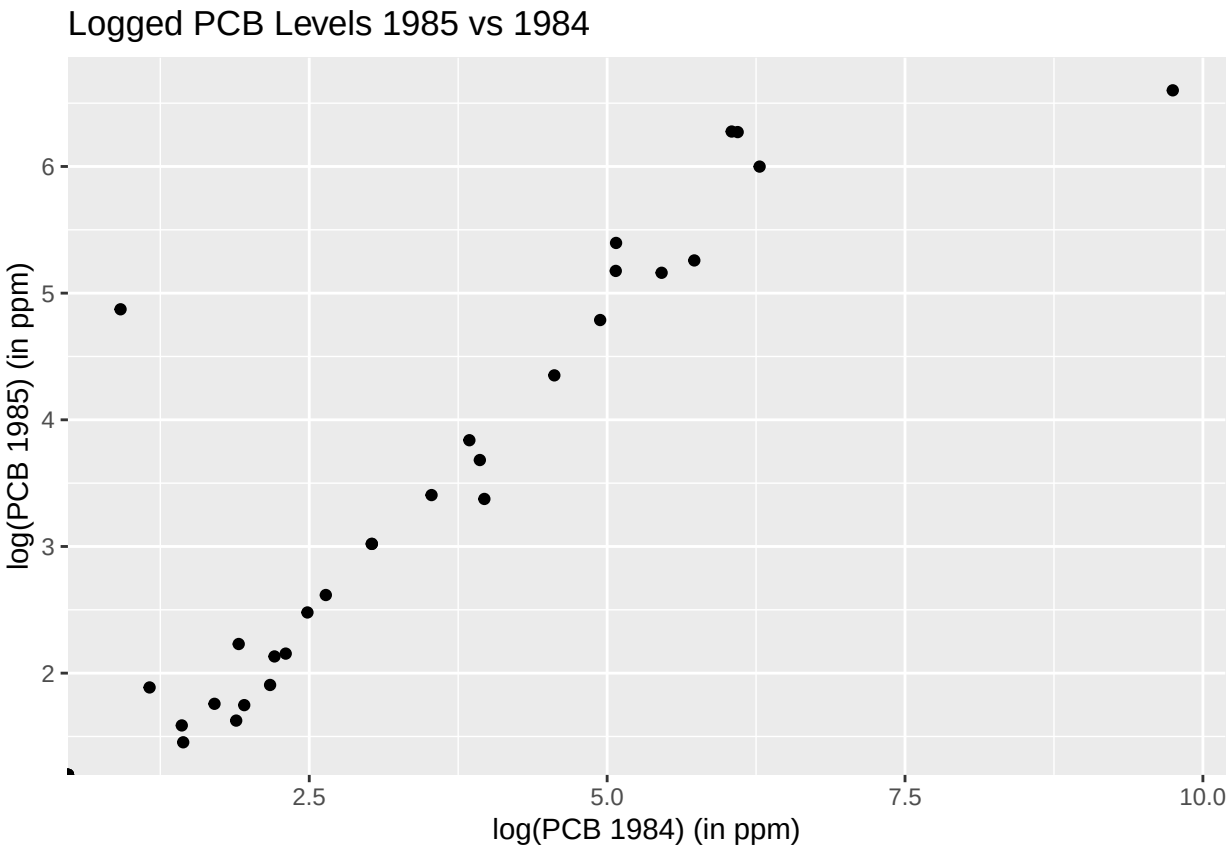
```
#removes outlier value (note that this data point was Boston Harbor)
pcb_new <- pcb[-4,]

#creates new scatterplot of pcb85 vs pcb84 - no outlier included
ggplot(data=pcb_new, aes(x = pcb84, y = pcb85)) +
  geom_point() +
    labs(title = "PCB Levels 1985 vs 1984 (No Outlier)", x = "PCB 1984 (in ppm)", y ="PCB 1985 (in ppm)
```




```
pcb_new %>%
  #creates SLR model on the new no outlier data
  lm(data = ., pcb85 ~ pcb84) %>%

  #augments model
  augment(.) %>%

  #creates and prints resid plot for model
  ggplot(., aes(x = pcb84, y = .resid)) +
    geom_point() +
    geom_hline(yintercept = 0, linetype = "dashed") +
    labs(title = "Pcb Resid Plot (No Outliers)", y = "Residuals", x = "Pcb84")
```




```
#creates scatterplot of pcb85 vs pcb84.
ggplot(data=pcb, aes(x = log(pcb84), y = log(pcb85))) +
  geom_point() +
  labs(title = "Logged PCB Levels 1985 vs 1984", x = "log(PCB 1984) (in ppm)", y ="log(PCB 1985) (in ppm
```


Logged PCB Levels 1985 vs 1984
log(PCB 1985) (in ppm)
6
5
4
3
2
2.5
5.0
7.5
10.0
log(PCB 1984) (in ppm)