# PSTAT 131/231 HW #1

*Lash Tan (231) and Jacobo Pereira-Pacheco (131)*

*4/19/2018*

*Appologies for not indenting, something in the RStudio environment has been breaking the output when doing any indenting*

```
algae <- read_table2("algaeBloom.txt", col_names =
                       c('season','size','speed','mxPH','mnO2','Cl','NO3','NH4',
                         'oPO4','PO4','Chla','a1','a2','a3','a4','a5','a6','a7'),
                     na="XXXXXXX")
```

```
## Parsed with column specification:
## cols(
##   season = col_character(),
##   size = col_character(),
##   speed = col_character(),
##   mxPH = col_double(),
##   mnO2 = col_double(),
##   Cl = col_double(),
##   NO3 = col_double(),
##   NH4 = col_double(),
##   oPO4 = col_double(),
##   PO4 = col_double(),
##   Chla = col_double(),
##   a1 = col_double(),
##   a2 = col_double(),
##   a3 = col_double(),
##   a4 = col_double(),
##   a5 = col_double(),
##   a6 = col_double(),
##   a7 = col_double()
## )
```

```
glimpse(algae)
```

```
## Observations: 200
## Variables: 18
## $ season <chr> "winter", "spring", "autumn", "spring", "autumn", "wint...
## $ size   <chr> "small", "small", "small", "small", "small", "small", "...
## $ speed  <chr> "medium", "medium", "medium", "medium", "medium", "high...
## $ mxPH   <dbl> 8.00, 8.35, 8.10, 8.07, 8.06, 8.25, 8.15, 8.05, 8.70, 7...
## $ mnO2   <dbl> 9.8, 8.0, 11.4, 4.8, 9.0, 13.1, 10.3, 10.6, 3.4, 9.9, 1...
## $ Cl     <dbl> 60.80, 57.75, 40.02, 77.36, 55.35, 65.75, 73.25, 59.07,...
## $ NO3    <dbl> 6.238, 1.288, 5.330, 2.302, 10.416, 9.248, 1.535, 4.990...
## $ NH4    <dbl> 578.00, 370.00, 346.67, 98.18, 233.70, 430.00, 110.00, ...
## $ oPO4   <dbl> 105.00, 428.75, 125.67, 61.18, 58.22, 18.25, 61.25, 44....
## $ PO4    <dbl> 170.00, 558.75, 187.06, 138.70, 97.58, 56.67, 111.75, 7...
## $ Chla   <dbl> 50.000, 1.300, 15.600, 1.400, 10.500, 28.400, 3.200, 6....
## $ a1     <dbl> 0.0, 1.4, 3.3, 3.1, 9.2, 15.1, 2.4, 18.2, 25.4, 17.0, 1...
## $ a2     <dbl> 0.0, 7.6, 53.6, 41.0, 2.9, 14.6, 1.2, 1.6, 5.4, 0.0, 0....
## $ a3     <dbl> 0.0, 4.8, 1.9, 18.9, 7.5, 1.4, 3.2, 0.0, 2.5, 0.0, 0.0,...
```

```
## $ a4     <dbl> 0.0, 1.9, 0.0, 0.0, 0.0, 0.0, 3.9, 0.0, 0.0, 2.9, 0.0, ...
## $ a5     <dbl> 34.2, 6.7, 0.0, 1.4, 7.5, 22.5, 5.8, 5.5, 0.0, 0.0, 1.2...
## $ a6     <dbl> 8.3, 0.0, 0.0, 0.0, 4.1, 12.6, 6.8, 8.7, 0.0, 0.0, 0.0,...
## $ a7     <dbl> 0.0, 2.1, 9.7, 1.4, 1.0, 2.9, 0.0, 0.0, 0.0, 1.7, 6.0, ...
```

***Question 1***

**a)**

There are 40 observations in Autumn, 53 observations in Spring, 45 observations in Summer, and 62 observations in Winter.

```
algae %>%
  group_by(season) %>%
  summarize(count_total = n())
```

```
## # A tibble: 4 x 2
##   season count_total
##    <chr>       <int>
## 1 autumn          40
## 2 spring          53
## 3 summer          45
## 4 winter          62
```

**b)**

Yes, there are several missing variables in the data set. Looking at solely the mean and variance of the two quanities for different chemicals, one notices that the magnitude is very different among different chemicals. For example, looking at $NO_3$ and $NH_4$, the averages are 3.28 and 501 respectively, and same magnitudinal difference apply to their variances. This can be attributed to values for these chemicals.

```
algae %>%
  summarize(mnO2_avg = mean(mnO2, na.rm=T), Cl_avg = mean(Cl, na.rm=T),     NO3_avg = mean(NO3, na.rm=T)
            NH4_avg = mean(NH4, na.rm=T),   oPO4_avg = mean(oPO4, na.rm=T), PO4_avg = mean(PO4, na.rm=T)
            Chla_avg = mean(Chla, na.rm=T),
            mnO2_var = var(mnO2, na.rm=T),  Cl_var = var(Cl, na.rm=T),      NO3_var = var(NO3, na.rm=T)
            NH4_var = var(NH4, na.rm=T),    oPO4_var = var(oPO4, na.rm=T),  PO4_var = var(PO4, na.rm=T)
            Chla_var = var(Chla, na.rm=T))
```

```
## # A tibble: 1 x 14
##   mnO2_avg Cl_avg NO3_avg NH4_avg oPO4_avg PO4_avg Chla_avg mnO2_var
##      <dbl>  <dbl>   <dbl>   <dbl>    <dbl>   <dbl>    <dbl>    <dbl>
## 1    9.118  43.64   3.282   501.3    73.59   137.9    13.97    5.718
## # ... with 6 more variables: Cl_var <dbl>, NO3_var <dbl>, NH4_var <dbl>,
## #   oPO4_var <dbl>, PO4_var <dbl>, Chla_var <dbl>
```

**c)**

It appears that for most chemicals the mean and median absolute difference (MAD) are fairly close to one another while the mean and variance can differ significantly.

```
algae %>%
  summarize(mnO2_med = median(mnO2, na.rm=T), Cl_med = median(Cl, na.rm=T),     NO3_med = median(NO3, na
            NH4_med = median(NH4, na.rm=T),   oPO4_med = median(oPO4, na.rm=T), PO4_med = median(PO4, na
            Chla_med = median(Chla, na.rm=T),
```

```
            mnO2_mad = mad(mnO2, na.rm=T),    Cl_mad = mad(Cl, na.rm=T),        NO3_mad = mad(NO3, na.rm
            NH4_mad = mad(NH4, na.rm=T),      oPO4_mad = mad(oPO4, na.rm=T),    PO4_mad = mad(PO4, na.rm
            Chla_mad = mad(Chla, na.rm=T))
```

```
## # A tibble: 1 x 14
##   mnO2_med Cl_med NO3_med NH4_med oPO4_med PO4_med Chla_med mnO2_mad
##      <dbl>  <dbl>   <dbl>   <dbl>    <dbl>   <dbl>    <dbl>    <dbl>
## 1      9.8  32.73   2.675   103.2    40.15   103.3    5.475    2.053
## # ... with 6 more variables: Cl_mad <dbl>, NO3_mad <dbl>, NH4_mad <dbl>,
## #   oPO4_mad <dbl>, PO4_mad <dbl>, Chla_mad <dbl>
```

```
algae_cts <- algae %>%
  summarize(mnO2_avg = mean(mnO2, na.rm=T), Cl_avg = mean(Cl, na.rm=T),     NO3_avg = mean(NO3, na.rm=T
            NH4_avg = mean(NH4, na.rm=T),   oPO4_avg = mean(oPO4, na.rm=T), PO4_avg = mean(PO4, na.rm=T
            Chla_avg = mean(Chla, na.rm=T),
            mnO2_var = var(mnO2, na.rm=T),  Cl_var = var(Cl, na.rm=T),      NO3_var = var(NO3, na.rm=T)
            NH4_var = var(NH4, na.rm=T),    oPO4_var = var(oPO4, na.rm=T),  PO4_var = var(PO4, na.rm=T)
            Chla_var = var(Chla, na.rm=T),
            mnO2_med = median(mnO2, na.rm=T), Cl_med = median(Cl, na.rm=T),     NO3_med = median(NO3, na
            NH4_med = median(NH4, na.rm=T),   oPO4_med = median(oPO4, na.rm=T), PO4_med = median(PO4, na
            Chla_med = median(Chla, na.rm=T),
            mnO2_mad = mad(mnO2, na.rm=T),    Cl_mad = mad(Cl, na.rm=T),        NO3_mad = mad(NO3, na.rm
            NH4_mad = mad(NH4, na.rm=T),      oPO4_mad = mad(oPO4, na.rm=T),    PO4_mad = mad(PO4, na.rm
            Chla_mad = mad(Chla, na.rm=T))

algae_cts %>%
  select(starts_with("mnO2"), starts_with("Cl"), starts_with("NO3")) %>%
  t()
```

```
##              [,1]
## mnO2_avg    9.118
## mnO2_var    5.718
## mnO2_med    9.800
## mnO2_mad    2.053
## Cl_avg     43.636
## Cl_var   2193.172
## Cl_med     32.730
## Cl_mad     33.250
## NO3_avg     3.282
## NO3_var    14.262
## NO3_med     2.675
## NO3_mad     2.172
```

```
algae_cts %>%
  select(starts_with("NH4"), starts_with("oPO4"), starts_with("PO4"), starts_with("Chla")) %>%
  round(4) %>%
  t()
```

```
##               [,1]
## NH4_avg  5.013e+02
## NH4_var  3.852e+06
## NH4_med  1.032e+02
## NH4_mad  1.116e+02
## oPO4_avg 7.359e+01
## oPO4_var 8.306e+03
```

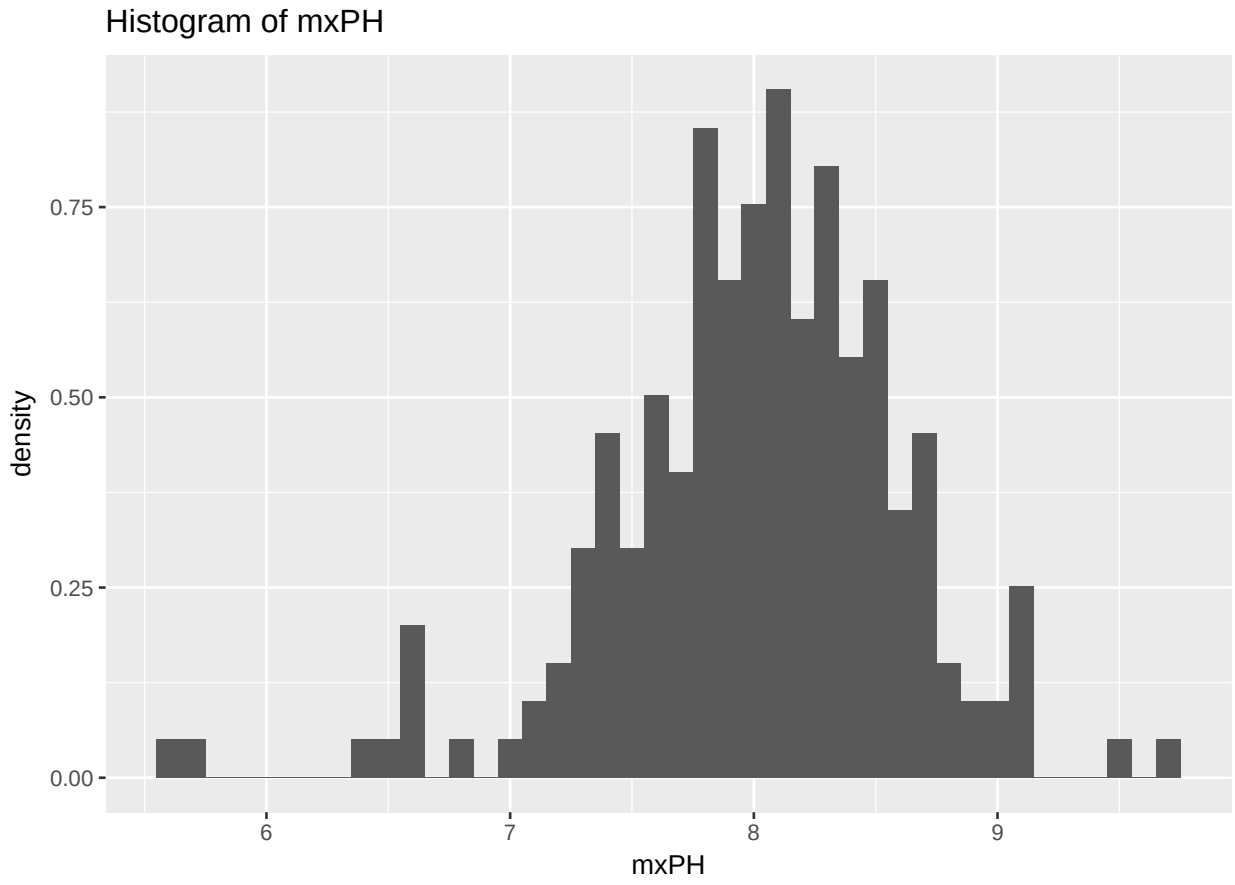
```
## oPO4_med 4.015e+01
## oPO4_mad 4.405e+01
## PO4_avg  1.379e+02
## PO4_var  1.664e+04
## PO4_med  1.033e+02
## PO4_mad  1.223e+02
## Chla_avg 1.397e+01
## Chla_var 4.201e+02
## Chla_med 5.475e+00
## Chla_mad 6.672e+00
```

***Question 2***

**a)**

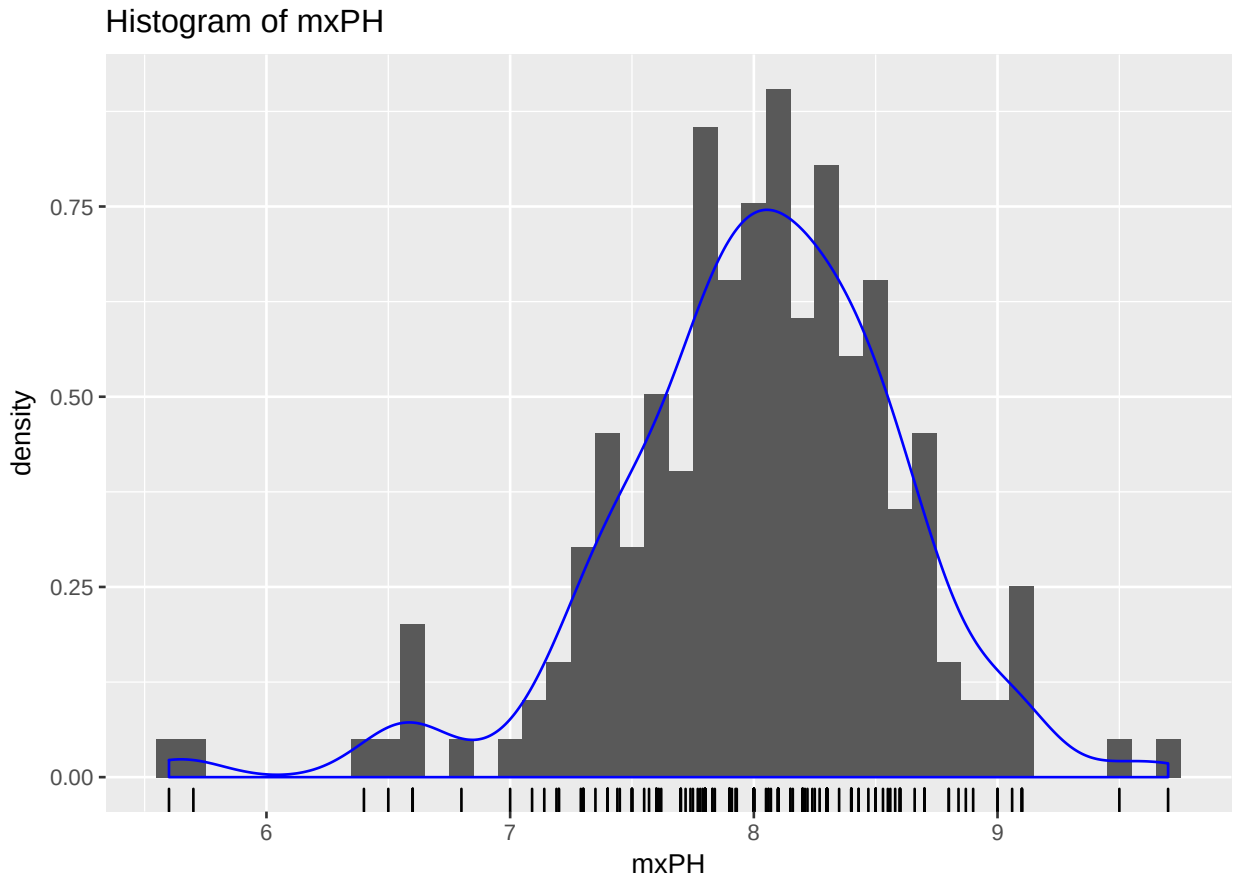
Yes, the distribution appears to be slightly negatively skewed with a good portion of the data tending to be on the right-side of the histogram. Different binwidths told different stories, but the binwidth we chose seemed to give the most accurate description of the data.

```
ggplot(algae) +
  geom_histogram(mapping = aes(x = mxPH, y = ..density..), binwidth = .1, na.rm = T) +
  labs(title = "Histogram of mxPH")
```



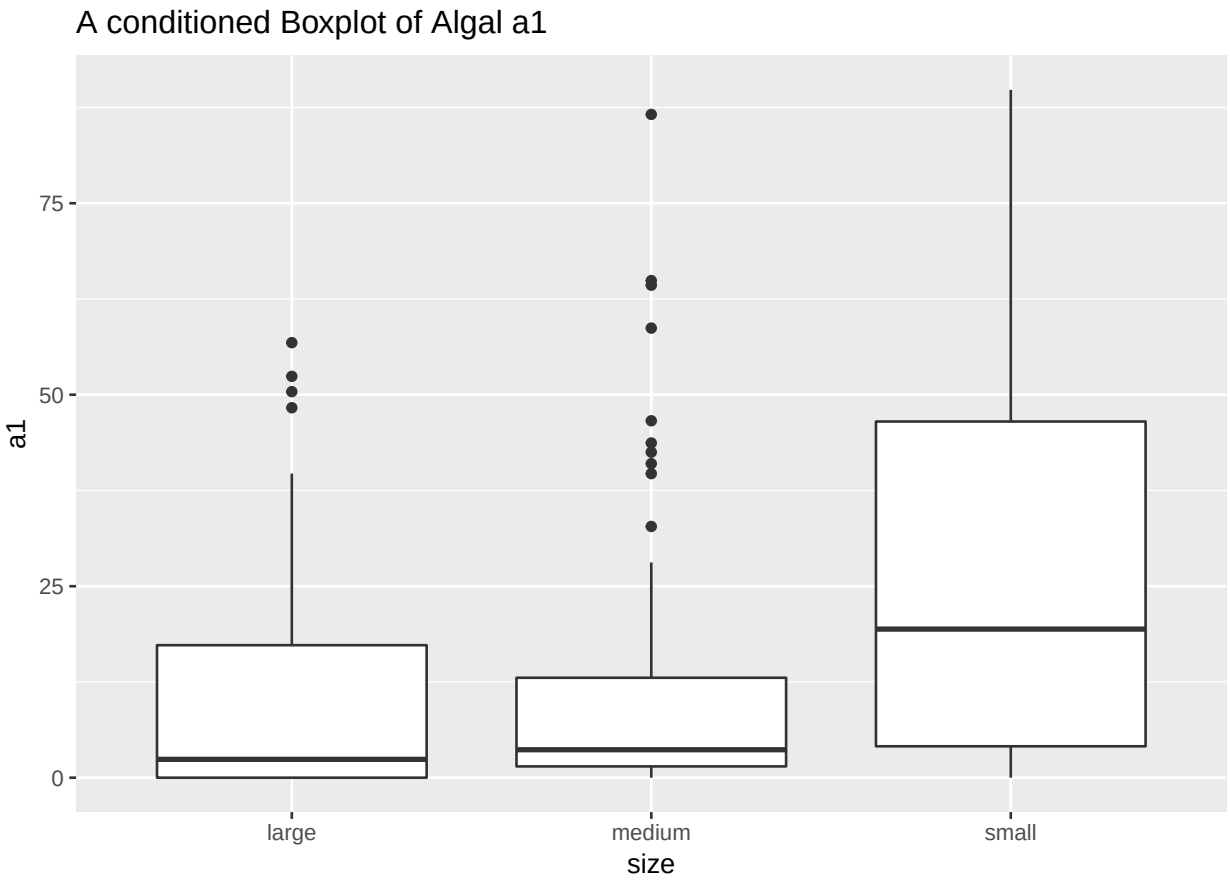

**b)**

```
ggplot(algae) +
  geom_histogram(mapping = aes(x = mxPH, y = ..density..), binwidth = .1, na.rm = T) +
  geom_density(mapping = aes(x = mxPH, y = ..density..), col = "blue") +
  geom_rug(mapping = aes(x = mxPH)) +
  labs(title = "Histogram of mxPH")
```




c)

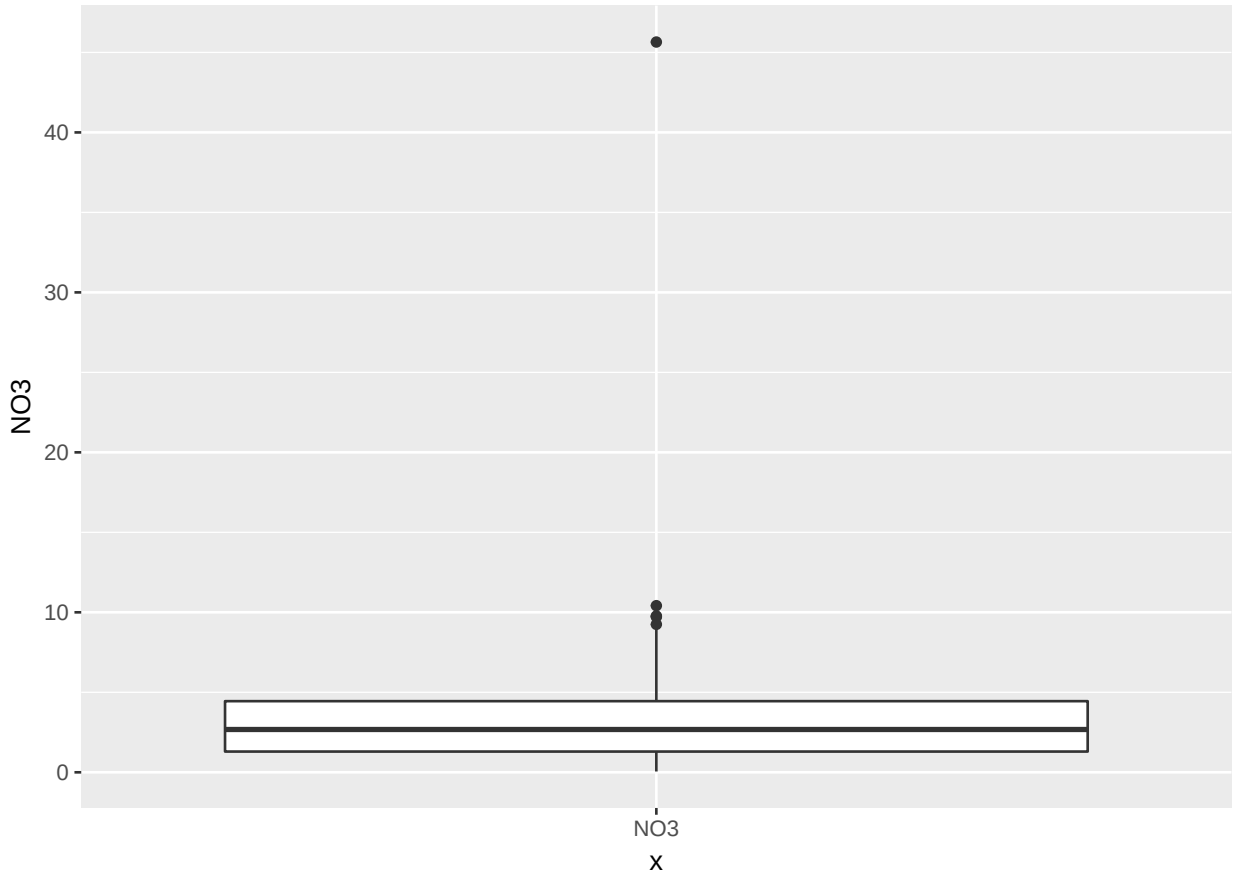
```
ggplot(algae) +
  geom_boxplot(aes(size, a1), na.rm=T) +
  labs(title = "A conditioned Boxplot of Algal a1")
```



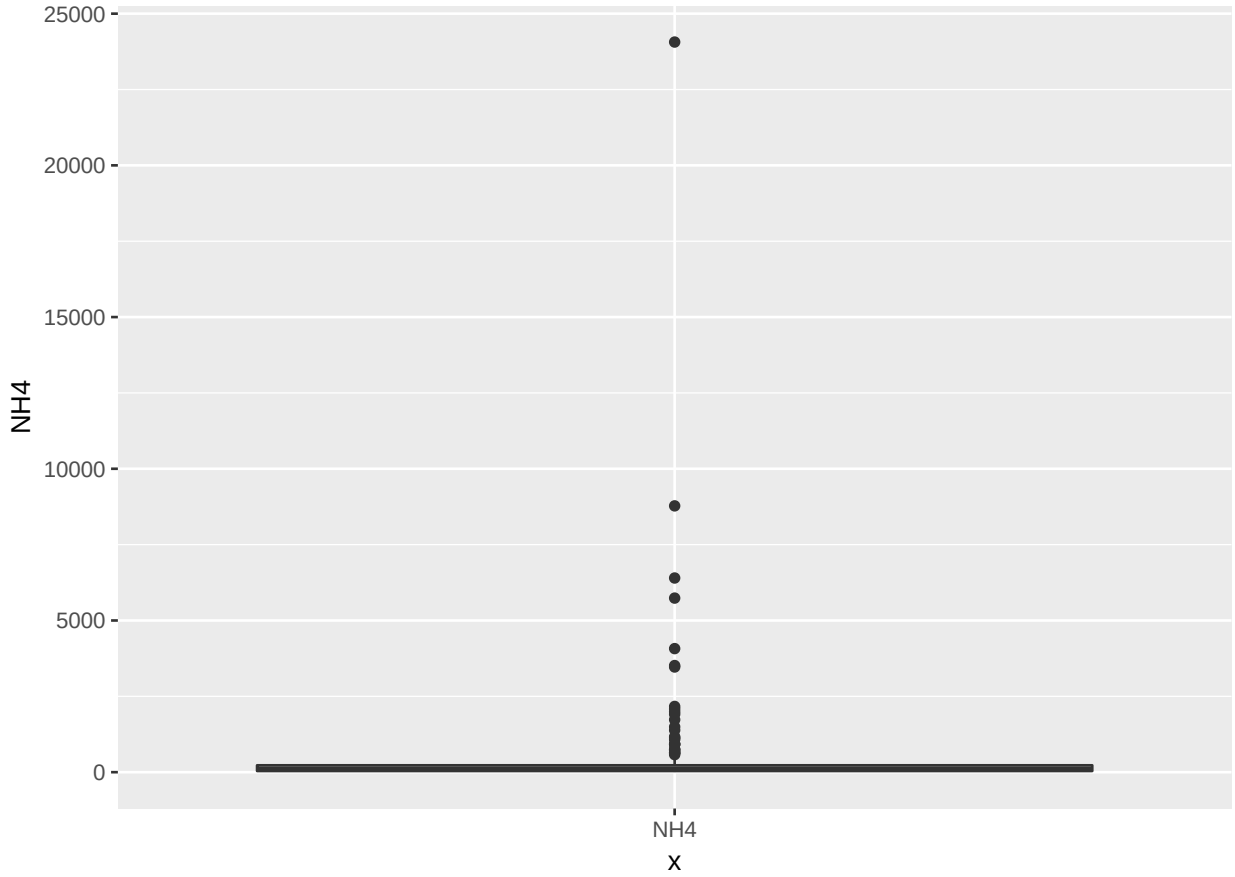

**d)**

Yes, outliers are present in both $NO_3$ and $NH_4$ in the positive direction of the boxplot. We would consider $NO_3$ to have 5 outliers and $NH_4$ to have 27. The number of outliers were determined by setting thresholds using interquantile ranges to set upper and lower bounds in the data, where observations above and below these thresholds would be considered outliers.

```
ggplot(algae) +
  geom_boxplot(aes(x = "NO3", y = NO3))
```




```
ggplot(algae) +
  geom_boxplot(aes(x = "NH4", y = NH4))
```




```
algae %>%
  filter(NO3 > (quantile(NO3, .75, na.rm=T) + 1.5 * IQR(NO3, na.rm=T)) ) %>%
  select(NO3) %>%
  arrange(desc(NO3))
```

```
## # A tibble: 5 x 1
##      NO3
##    <dbl>
## 1 45.650
## 2 10.416
## 3  9.773
## 4  9.715
## 5  9.248
```

```
algae %>%
  filter(NH4 > (quantile(NH4, .75, na.rm=T) + 1.5 * IQR(NH4, na.rm=T)) ) %>%
  select(NH4) %>%
  arrange(desc(NH4))
```

```
## # A tibble: 27 x 1
##      NH4
##    <dbl>
##  1 24064
##  2  8778
##  3  6400
##  4  5738
```

```
##  5  4073
##  6  3515
##  7  3467
##  8  2167
##  9  2083
## 10  1990
## # ... with 17 more rows
```

**e)**

It's clear that the measurements for $NH_4$ are on a much larger scale compared to the measurements of $NH_3$. However the trends for each measurements between the chemicals appear to be similar, i.e, for both chemicals the median is slightly larger than the MAD. Seeing that the variancee for $NH_4$ are exceedingly large, we would conclude that the median and MAD are more robust to outliers.

```
algae_cts %>%
  select(starts_with("NO3") , starts_with("NH4"))
```

```
## # A tibble: 1 x 8
##   NO3_avg NO3_var NO3_med NO3_mad NH4_avg NH4_var NH4_med NH4_mad
##     <dbl>   <dbl>   <dbl>   <dbl>   <dbl>   <dbl>   <dbl>   <dbl>
## 1   3.282   14.26   2.675   2.172   501.3 3851585   103.2   111.6
```

***Question 3***

**a)**

There are 33 observations with missing values with 1 missing value in $mxPH$, 2 missing values in $mnO_2$, 10 missing values in $Cl$, 2 missing values in $NO_3$, 2 missing values in $NH_4$, 2 missing values in $oPO_4$, 2 missing values in $PO_4$, and 12 missing values in $Chla$.

```
algae %>% is.na() %>% sum()
```

```
## [1] 33
```

```
algae %>% is.na()%>% colSums()
```

```
## season   size  speed   mxPH   mnO2     Cl    NO3    NH4   oPO4    PO4
##      0      0      0      1      2     10      2      2      2      2
##   Chla     a1     a2     a3     a4     a5     a6     a7
##     12      0      0      0      0      0      0      0
```

**b)**

There are 184 observations in algae.del.

```
(algae.del <- algae %>%
  filter(complete.cases(algae)))
```

```
## # A tibble: 184 x 18
##    season  size  speed  mxPH  mnO2    Cl    NO3    NH4   oPO4    PO4
##     <chr> <chr>  <chr> <dbl> <dbl> <dbl>  <dbl>  <dbl>  <dbl>  <dbl>
##  1 winter small medium  8.00   9.8 60.80  6.238 578.00 105.00 170.00
##  2 spring small medium  8.35   8.0 57.75  1.288 370.00 428.75 558.75
##  3 autumn small medium  8.10  11.4 40.02  5.330 346.67 125.67 187.06
##  4 spring small medium  8.07   4.8 77.36  2.302  98.18  61.18 138.70
```

```
##  5 autumn small medium  8.06   9.0 55.35 10.416 233.70  58.22  97.58
##  6 winter small   high  8.25  13.1 65.75  9.248 430.00  18.25  56.67
##  7 summer small   high  8.15  10.3 73.25  1.535 110.00  61.25 111.75
##  8 autumn small   high  8.05  10.6 59.07  4.990 205.67  44.67  77.43
##  9 winter small medium  8.70   3.4 21.95  0.886 102.75  36.30  71.00
## 10 winter small   high  7.93   9.9  8.00  1.390   5.80  27.25  46.60
## # ... with 174 more rows, and 8 more variables: Chla <dbl>, a1 <dbl>,
## #   a2 <dbl>, a3 <dbl>, a4 <dbl>, a5 <dbl>, a6 <dbl>, a7 <dbl>
```

**c)**

```
algae.med <- algae %>%
  mutate_at(vars(mxPH:Chla), funs(ifelse(is.na(.), median(., na.rm=T), .)))
algae.med %>%
  select_at(vars(mnO2:Chla)) %>%
  slice(c(48, 62, 199))
```

```
## # A tibble: 3 x 7
##    mnO2    Cl   NO3   NH4  oPO4   PO4  Chla
##   <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>
## 1  12.6  9.00 0.230  10.0  5.00   6.0 1.100
## 2   9.8 32.73 2.675 103.2 40.15  14.0 5.475
## 3   7.6 32.73 2.675 103.2 40.15 103.3 5.475
```

**d)**

We obtain 48.06929 for the $28^{th}$ observation

```
algae %>%
  select_at(vars(mxPH:Chla)) %>%
  cor(use="complete.obs")
```

```
##          mxPH     mnO2       Cl     NO3      NH4     oPO4     PO4    Chla
## mxPH  1.00000 -0.10269  0.14710 -0.1721 -0.15430  0.09023  0.1013  0.4318
## mnO2 -0.10269  1.00000 -0.26325  0.1179 -0.07827 -0.39375 -0.4640 -0.1312
## Cl    0.14710 -0.26325  1.00000  0.2110  0.06598  0.37926  0.4452  0.1430
## NO3  -0.17213  0.11791  0.21096  1.0000  0.72468  0.13301  0.1570  0.1455
## NH4  -0.15430 -0.07827  0.06598  0.7247  1.00000  0.21931  0.1994  0.0912
## oPO4  0.09023 -0.39375  0.37926  0.1330  0.21931  1.00000  0.9120  0.1069
## PO4   0.10133 -0.46396  0.44519  0.1570  0.19940  0.91196  1.0000  0.2485
## Chla  0.43182 -0.13122  0.14296  0.1455  0.09120  0.10691  0.2485  1.0000
```

```
algae %>%
  select("oPO4") %>%
  slice(28)
```

```
## # A tibble: 1 x 1
##    oPO4
##   <dbl>
## 1     4
```

```
predict(lm(PO4~oPO4, algae), data.frame(oPO4 = 4))
```

```
##     1
## 48.07
```

**e)**

Incorrect conclusions from only the observed data may occur if the dataset is too small. In particular, in some scenarios it is most useful to understand as to why some missing values exist. In particular, it is important to recall the example with the airplanes that were shot down, where Abraham Wald recognized that planes should be reinforced where missing data occurred and those planes were not found. This is true in a universal scenario, instead of simply thinking of techniques to fill in missing values, it may be more useful to understand why those values are missing to begin with. This is the essence of survivorship bias.

### *Question 4*

**a)**

```
(algae.chk <- algae.med %>%
  mutate(chk = sample(cut(seq(1,200,1),5, label=F))))
```

```
## # A tibble: 200 x 19
##    season  size  speed  mxPH  mnO2    Cl    NO3    NH4   oPO4    PO4
##     <chr> <chr>  <chr> <dbl> <dbl> <dbl>  <dbl>  <dbl>  <dbl>  <dbl>
##  1 winter small medium  8.00   9.8 60.80  6.238 578.00 105.00 170.00
##  2 spring small medium  8.35   8.0 57.75  1.288 370.00 428.75 558.75
##  3 autumn small medium  8.10  11.4 40.02  5.330 346.67 125.67 187.06
##  4 spring small medium  8.07   4.8 77.36  2.302  98.18  61.18 138.70
##  5 autumn small medium  8.06   9.0 55.35 10.416 233.70  58.22  97.58
##  6 winter small   high  8.25  13.1 65.75  9.248 430.00  18.25  56.67
##  7 summer small   high  8.15  10.3 73.25  1.535 110.00  61.25 111.75
##  8 autumn small   high  8.05  10.6 59.07  4.990 205.67  44.67  77.43
##  9 winter small medium  8.70   3.4 21.95  0.886 102.75  36.30  71.00
## 10 winter small   high  7.93   9.9  8.00  1.390   5.80  27.25  46.60
## # ... with 190 more rows, and 9 more variables: Chla <dbl>, a1 <dbl>,
## #   a2 <dbl>, a3 <dbl>, a4 <dbl>, a5 <dbl>, a6 <dbl>, a7 <dbl>, chk <int>
```

**b)**

```
do.chunk <- function(chunkid, chunkdef, dat){ # function argument
  train = (chunkdef != chunkid)
  Xtr = dat[train,1:11] # get training set
  Ytr = dat[train,12] # get true response values in trainig set
  Xvl = dat[!train,1:11] # get validation set
  Yvl = dat[!train,12] # get true response values in validation set
  lm.a1 <- lm(a1~., data = dat[train,1:12])

  predYtr = predict(lm.a1) # predict training values
  predYvl = predict(lm.a1,Xvl) # predict validation values
  data.frame(fold = chunkid,
             train.error = mean((predYtr - Ytr)^2), # compute and store training error
             val.error = mean((predYvl - Yvl)^2)) # compute and store test error
}

algae.chk %>%
    lapply(c(1:5), do.chunk, chunkdef = .$chk, dat = .)
```

```
## [[1]]
##   fold train.error val.error
## 1    1       304.9     235.9
```

```
##
## [[2]]
##   fold train.error val.error
## 1    2       247.5     671.4
##
## [[3]]
##   fold train.error val.error
## 1    3       262.6     415.5
##
## [[4]]
##   fold train.error val.error
## 1    4       277.1     364.5
##
## [[5]]
##   fold train.error val.error
## 1    5       299.9     263.2
```

***Question 5***

**a)**

Yes,this is expected as the "true" test error is is around the average of the estimated test error from question 4.

```
algae.Test <- read_table2('algaeTest.txt',
                          col_names=c('season','size','speed','mxPH','mnO2','Cl','NO3',
                                      'NH4','oPO4','PO4','Chla','a1'),
                          na=c('XXXXXXX'))
```

```
## Parsed with column specification:
## cols(
##   season = col_character(),
##   size = col_character(),
##   speed = col_character(),
##   mxPH = col_double(),
##   mnO2 = col_double(),
##   Cl = col_double(),
##   NO3 = col_double(),
##   NH4 = col_double(),
##   oPO4 = col_double(),
##   PO4 = col_double(),
##   Chla = col_double(),
##   a1 = col_double()
## )
```

```
algae.merged <- rbind(
  algae %>%
    select(season:a1) %>%
    mutate(chk = 1),
  algae.Test %>%
    mutate(chk = 2))

algae.merged %>%
  do.chunk(2, .$chk, .)
```

```
##   fold train.error val.error
## 1    2       457.5     249.2
```

***Question 6***

**a)**

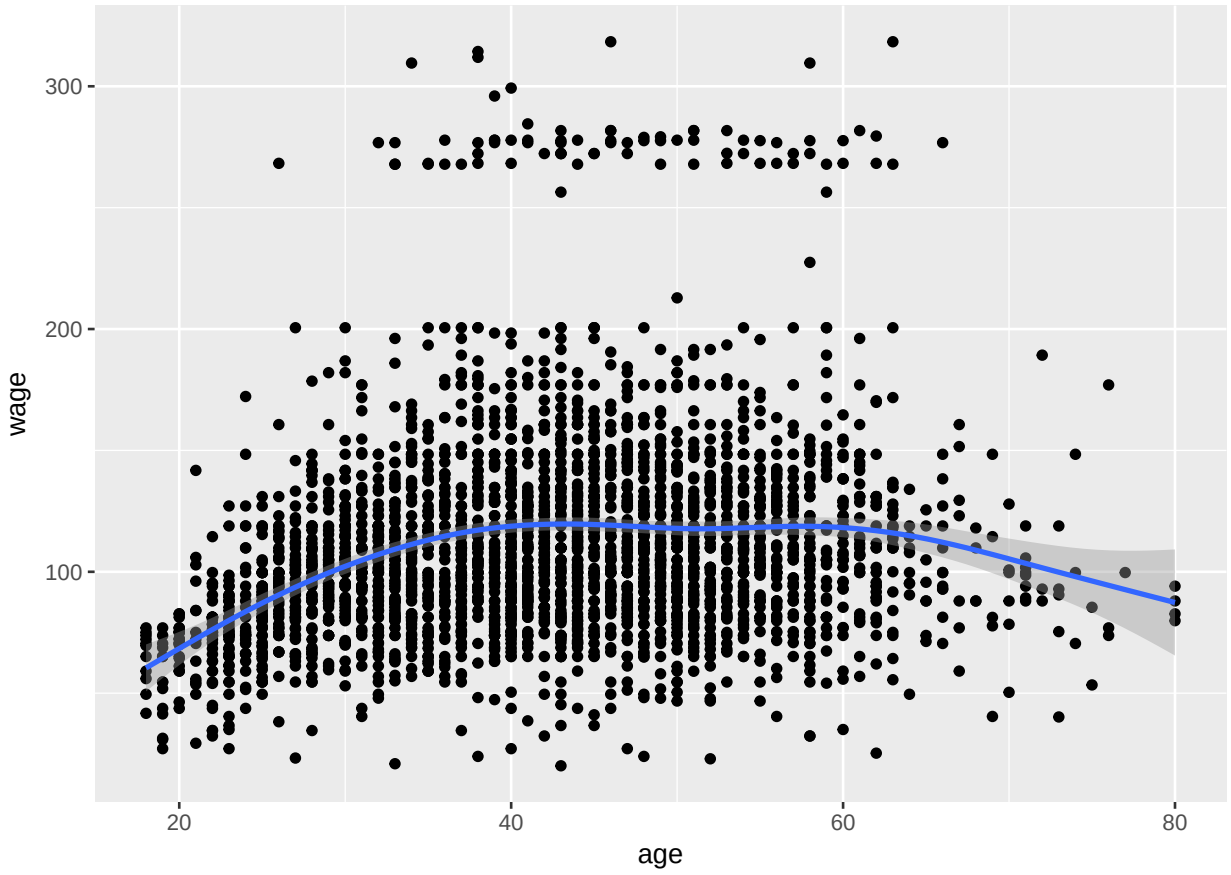
The plot shows that wages at the age extremes (youngers and older) tend to earn less/have a lower ceiling, which is to be expected. The prime working years have more people earning higher salaries.

```
head(Wage)
```

```
##        year age           maritl     race       education
## 231655 2006  18 1. Never Married 1. White    1. < HS Grad
## 86582  2004  24 1. Never Married 1. White 4. College Grad
## 161300 2003  45       2. Married 1. White 3. Some College
## 155159 2003  43       2. Married 3. Asian 4. College Grad
## 11443  2005  50      4. Divorced 1. White      2. HS Grad
## 376662 2008  54       2. Married 1. White 4. College Grad
##                      region       jobclass         health health_ins logwage
## 231655 2. Middle Atlantic  1. Industrial      1. <=Good      2. No   4.318
## 86582  2. Middle Atlantic 2. Information 2. >=Very Good      2. No   4.255
## 161300 2. Middle Atlantic  1. Industrial      1. <=Good     1. Yes   4.875
## 155159 2. Middle Atlantic 2. Information 2. >=Very Good     1. Yes   5.041
## 11443  2. Middle Atlantic 2. Information      1. <=Good     1. Yes   4.318
## 376662 2. Middle Atlantic 2. Information 2. >=Very Good     1. Yes   4.845
##          wage
## 231655  75.04
## 86582   70.48
## 161300 130.98
## 155159 154.69
## 11443   75.04
## 376662 127.12
```

```
ggplot(Wage, mapping = aes(x=age, y=wage)) +
  geom_point() +
  geom_smooth()
```

```
## `geom_smooth()` using method = 'gam'
```




b)

```
library(plyr)
```

```
## ------------------------------------------------------------------------------
## You have loaded plyr after dplyr - this is likely to cause problems.
## If you need functions from both plyr and dplyr, please load plyr first, then dplyr:
## library(plyr); library(dplyr)
## ------------------------------------------------------------------------------
##
## Attaching package: 'plyr'
## The following objects are masked from 'package:dplyr':
##
##     arrange, count, desc, failwith, id, mutate, rename, summarise,
##     summarize
## The following object is masked from 'package:purrr':
##
##     compact
```

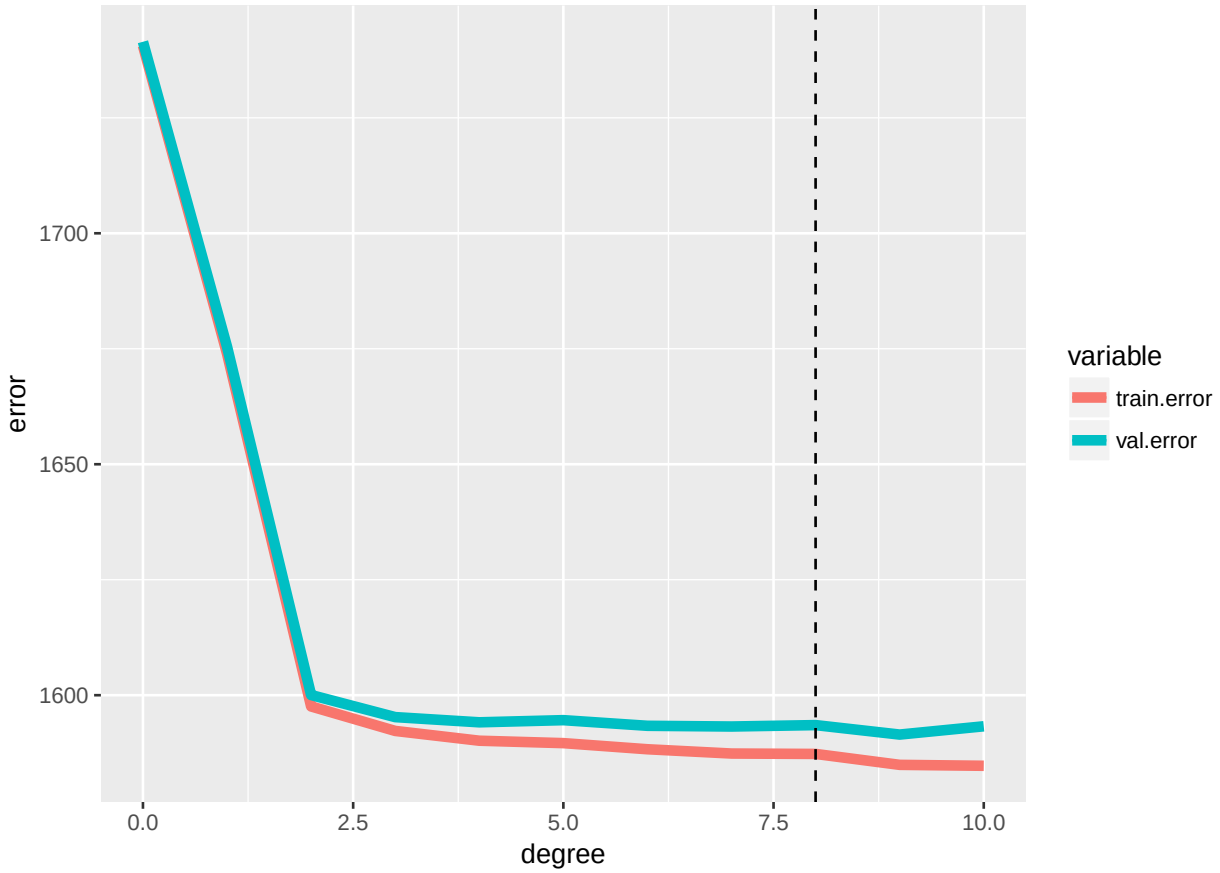
```
wage.chk <- Wage %>%
  mutate(chk = sample(cut(1:3000,5, label=F))) %>%
  cbind(.,data.frame(poly(Wage$age, 10, raw=F))) %>%
  select_at(vars(X1:X10, age, wage, chk))
```

```
do.chunky <- function(chunkid, chunkdef, dat, p){
  train.dat = dat %>%
    filter(dat$chk != chunkid)
  train = (chunkdef != chunkid)

  if (p == 0) lm.wage <- lm(wage~1, data = train.dat)
  else lm.wage <- lm(wage~., data = train.dat[,c(1:p,12)])

  Ytr = dat[train,12] # get true response values in training set
  if(p == 0){
    Xvl <- dat %>%
      filter(chunkdef == chunkid) %>%
      select(age)
  }
  else {
    Xvl <- dat %>%
      filter(chunkdef == chunkid) %>%
      select(1:p)
  }

  Yvl = dat[!train,12] # get true response values in validation set
  predYtr = predict(lm.wage) # predict training values
  predYvl = predict(lm.wage,Xvl) # predict validation values
  data.frame(degree = p,
             train.error = mean((predYtr - Ytr)^2), # compute and store training error
             val.error = mean((predYvl - Yvl)^2)) # compute and store test error

}
df.bind <- NULL
i <- 0
for (i in 0:10){
  df.bind <- rbind(df.bind, ldply(1:5, do.chunky, chunkdef = wage.chk$chk, dat = wage.chk, p =i))
}

err.avgs <- df.bind %>%
  group_by(degree) %>%
  summarize_all(mean)
```

c)

```
cv = 8
melted.wage <- melt(err.avgs, id.vars='degree', value.name = 'error')

ggplot(melted.wage, aes(x=degree, y=error, color=variable)) +
  geom_line(aes(group=interaction(variable,degree))) +
  stat_summary(aes(group=variable), fun.y="mean", geom='line', size=2) +
  geom_vline(aes(xintercept=cv), linetype='dashed')
```

```
## geom_path: Each group consists of only one observation. Do you need to
## adjust the group aesthetic?
```




The training error and test error quickly decline as the degree of age increases. We expect the training error to be monotonic as the degree increases, but we notice that the training error starts increasing past the 8th degree. We may choose this as our model, but we also notice that there is very little difference between the 8th degree errors and the 4th or 5th degree, so we may want to simply choose the 4th degree for simplicity.

# SOLUTION TO HOMEWORK 1

Lash Tan
PSTAT 231 – Spring '18

---

7. The bias-variance tradeoff. Prove that the mean squared error can be decomposed into the variance plus bias squared.

**Solution:**

$$MSE_{\hat{\theta}} = \mathbb{E}[(\hat{\theta} - \theta)^2] = Var(\hat{\theta}) + Bias(\hat{\theta})^2$$

Note that $\text{Bias}(\hat{\theta}) = \mathbb{E}[\hat{\theta}] - \theta$, $\mathbb{E}[\theta] = \theta$ as $\theta$ is a constant, and linearity and independence of expected value properties hold. Then,

$$\mathbb{E}[(\hat{\theta} - \theta)^2] = \mathbb{E}[\hat{\theta}^2 - 2\hat{\theta}\theta + \theta^2] = \mathbb{E}[\hat{\theta}^2] - 2\theta\mathbb{E}[\hat{\theta}] + \mathbb{E}[\theta^2]$$

$$= \mathbb{E}[\hat{\theta}^2] - \mathbb{E}[\hat{\theta}]^2 + \mathbb{E}[\hat{\theta}]^2 - 2\theta\mathbb{E}[\hat{\theta}] + \theta^2$$

$$= (\mathbb{E}[\hat{\theta}^2] - \mathbb{E}[\hat{\theta}]^2) + (\mathbb{E}[\hat{\theta}] - \theta)^2$$

$$= Var(\hat{\theta}) + [Bias(\hat{\theta})]^2$$

8. Show that the following measures are distance metrics by showing the listed properties hold:

- *Positivity*:
  - $d(x, y) \geq 0$
  - $d(x, y) = 0$ only if $x = y$
- *Symmetry*:
  - $d(x, y) = d(y, x)$ for all $x$ and $y$
- $Triangle Inequality$:
  - $d(x, z) \leq d(x, y) + d(y, z)$ for $x$, $y$, and $z$

(a) $d(x, y) = ||x - y||_2$

**Solution:** For parts (a) and (b), let $a \in x, b \in y$, and $c \in z$ be arbitrary for $x, y, z \in \mathbb{R}$

$$||x - y||_2 = (\sum_{j=1}^{n} |x_j - y_j|^p)^{1/p} = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2}$$

- Positivity: Let $a = b$. Then,

$$||a - b||_2 = ||a - a||_2 = (\sum_{j=1}^{n} |x_j - x_j|^2)^{1/2}$$

$$= (\sum_{j=1}^{n} |0|^2)^{1/2} = 0$$

Now, suppose $a \neq b$, or $a - b \neq 0$. Then,

$$||a-b||_2 = \sqrt{(x_1-y_1)^2 + (x_2-y_2)^2}$$

As $a \neq b$, $x_1 - y_1 \neq 0$ and $x_2 - y_2 \neq 0$, so $(x_1-y_1)^2 > 0$ and $(x_2-y_2)^2 > 0$, and the square root of a positive number is a positive number, and therefore $||a-b||_2 = ||x-y||_2 > 0$ for $x \neq y$. Combining these two, we satisfy positivity.

- Symmetry: $d(x,y) = d(y,x)$ or $||x-y||_2 = ||y-x||_2 \; \forall x, y$ Note that if $a - b \geq 0$ or $a \geq b$ then $|a-b| = a-b$ and if $a - b \leq 0$ or $a \leq b$ then $|a-b| = b-a$. And, if $b - a \geq 0$ or $a \leq b$ then $|b-a| = b-a$ and if $b-a \leq 0$ or $a \geq b$, then $|b-a| = a-b$. So, if $a \geq b, |a-b| = a-b$ and $|b-a| = a-b$, and if $a \leq b, |a-b| = b-a$ and $|b-a| = b-a$. So, $\forall a, b, |a-b| = |b-a|$. Therefore, $\sqrt{(x_1-y_1)^2 + (x_2-y_2)^2} = \sqrt{(y_1-x_1)^2 + (y_2-x_2)^2}$, and so we satisfy the symmetric property.

- Triangle Inequality: Using Minkowski inequality where:

$$(\sum_{k=1}^{n} |x_k + y_k|^p)^{1/p} \leq (\sum_{k=1}^{n} |x_k|^p)^{1/p} + (\sum_{k=1}^{n} |y_k|^p)^{1/p}$$

Applying this,

$$(\sum_{j=1}^{2} |x_i - y_i|^2)^{1/2} + (\sum_{j=1}^{2} |y_i - z_i|^2)^{1/2} \geq (\sum_{j=1}^{2} (|x_i - y_i| + |y_i - z_i|)^2)^{1/2}$$

$$\geq (\sum_{j=1}^{2} |x_i - z_i|)^2)^{1/2} = ||x_i - z_i||_2$$

(b) $d(x,y) = ||x-y||_\infty$

**Solution:**

$$||x-y||_\infty = \max_{1 \leq i \leq n} |x_i - y_i|$$

- Positivity: Let $a = b$. Then, $a-b = a-a = 0$, so $|x_i - y_i| = 0$, and so $\max_{1 \leq j \leq n} |x_i - y_i| = 0$. Now, suppose $a \neq b$. Then, $a - b \neq 0$, so $|a-b| \neq 0$, and since $|a-b| \geq 0$ and $|a-b| \neq 0, |a-b| > 0$. Therefore, $\max_{1 \leq j \leq n} = d(x,y) > 0$ if $x \neq y$, and $d(x,y) = 0$ if $x = y$, so $d(x,y) \geq 0$.

- Symmetry: As shown in (a), $\forall a, b$, $|a-b| = |b-a|$. Therefore, $\max_{1 \leq j \leq n} |x_i - y_i| = \max_{1 \leq j \leq n} |y_i - x_i|$, and $d(x,y) = d(y,x)$.

- Triangle Inequality: Note that $|a-c| = |a-b+b-c|$. The triangle inequality for real numbers says that $|x+y| \leq |x| + |y| \forall x, y \in \mathbb{R}$. So, $|a-b+b-c| \leq |a-b| + |b-c|$. Since $|a-b| \leq \max_{1 \leq j \leq n} |x_i - y_i|$ and $|b-c| \leq \max_{1 \leq j \leq n} |y_i - z_i|, |a-b| + |b-c| \leq \max_{1 \leq j \leq n} |x_i - y_i| + \max_{1 \leq j \leq n} |y_i - z_i|$, or $|a-c| \leq |a-b| + |b-c| \leq \max_{1 \leq j \leq n} |x_i - y_i| + \max_{1 \leq j \leq n} |y_i - z_i|$, so $d(x,z) \leq d(x,y) + d(y,z)$.