Homework 1

Julissa Duenas

1/22/2021

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

```
##
## -- 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()
## )
```

```r
glimpse(algae)
```

```
## Rows: 200
## Columns: 18
## $ season <chr> "winter", "spring", "autumn", "spring", "autumn", "winter", ...
## $ size   <chr> "small", "small", "small", "small", "small", "small", "small...
## $ speed  <chr> "medium", "medium", "medium", "medium", "medium", "high", "h...
## $ mxPH   <dbl> 8.00, 8.35, 8.10, 8.07, 8.06, 8.25, 8.15, 8.05, 8.70, 7.93, ...
## $ mnO2   <dbl> 9.8, 8.0, 11.4, 4.8, 9.0, 13.1, 10.3, 10.6, 3.4, 9.9, 10.2, ...
## $ Cl     <dbl> 60.800, 57.750, 40.020, 77.364, 55.350, 65.750, 73.250, 59.0...
## $ NO3    <dbl> 6.238, 1.288, 5.330, 2.302, 10.416, 9.248, 1.535, 4.990, 0.8...
## $ NH4    <dbl> 578.000, 370.000, 346.667, 98.182, 233.700, 430.000, 110.000...
## $ oPO4   <dbl> 105.000, 428.750, 125.667, 61.182, 58.222, 18.250, 61.250, 4...
## $ PO4    <dbl> 170.000, 558.750, 187.057, 138.700, 97.580, 56.667, 111.750,...
## $ Chla   <dbl> 50.000, 1.300, 15.600, 1.400, 10.500, 28.400, 3.200, 6.900, ...
## $ a1     <dbl> 0.0, 1.4, 3.3, 3.1, 9.2, 15.1, 2.4, 18.2, 25.4, 17.0, 16.6, ...
## $ a2     <dbl> 0.0, 7.6, 53.6, 41.0, 2.9, 14.6, 1.2, 1.6, 5.4, 0.0, 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, 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, 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, 0.0...
## $ a6     <dbl> 8.3, 0.0, 0.0, 0.0, 4.1, 12.6, 6.8, 8.7, 0.0, 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, 1.5, ...
```

1a)

```
algae %>%
  group_by(season) %>%
  dplyr::summarise(obs=n(),na.rm=TRUE)
```

```
## # A tibble: 4 x 3
##   season   obs na.rm
## * <chr>  <int> <lgl>
## 1 autumn    40 TRUE
## 2 spring    53 TRUE
## 3 summer    45 TRUE
## 4 winter    62 TRUE
```

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

1b)

```
#is.na(algae)
Chemicals <- algae%>%select(mxPH:Chla)
Chemicals_mean <- Chemicals%>%summarise_all(mean,na.rm=TRUE)
Chemicals_var <- Chemicals%>%summarise_all(var,na.rm=TRUE)
print(Chemicals_mean)
```

```
## # A tibble: 1 x 8
##    mxPH  mnO2    Cl   NO3   NH4  oPO4   PO4  Chla
##   <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>
## 1  8.01  9.12  43.6  3.28  501.  73.6  138.  14.0
```

```
print(Chemicals_var)
```

```
## # A tibble: 1 x 8
##    mxPH  mnO2    Cl   NO3      NH4  oPO4    PO4  Chla
##   <dbl> <dbl> <dbl> <dbl>    <dbl> <dbl>  <dbl> <dbl>
## 1 0.358  5.72 2193.  14.3 3851585. 8306. 16639.  420.
```

The line is.na(algae) shows us that yes, there are missing values. I noticed that a higher mean warrants a higher variance execpt in the case of mxPH and mnO2. NH4 has the greatest variance as well and the greatest magnitude from its mean. NO3 has the smallest variance as well as the smallest magnitude from its mean

1c)

```
Chemicals_med <- Chemicals%>%summarise_all(median,na.rm=TRUE)
Chemicals_MAD <- Chemicals%>%summarise_all(mad,na.rm=TRUE)
print(Chemicals_med)
```

```
## # A tibble: 1 x 8
##    mxPH  mnO2    Cl   NO3   NH4  oPO4   PO4  Chla
##   <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>
## 1  8.06   9.8  32.7  2.68  103.  40.2  103.  5.48
```

```
print(Chemicals_MAD)
```

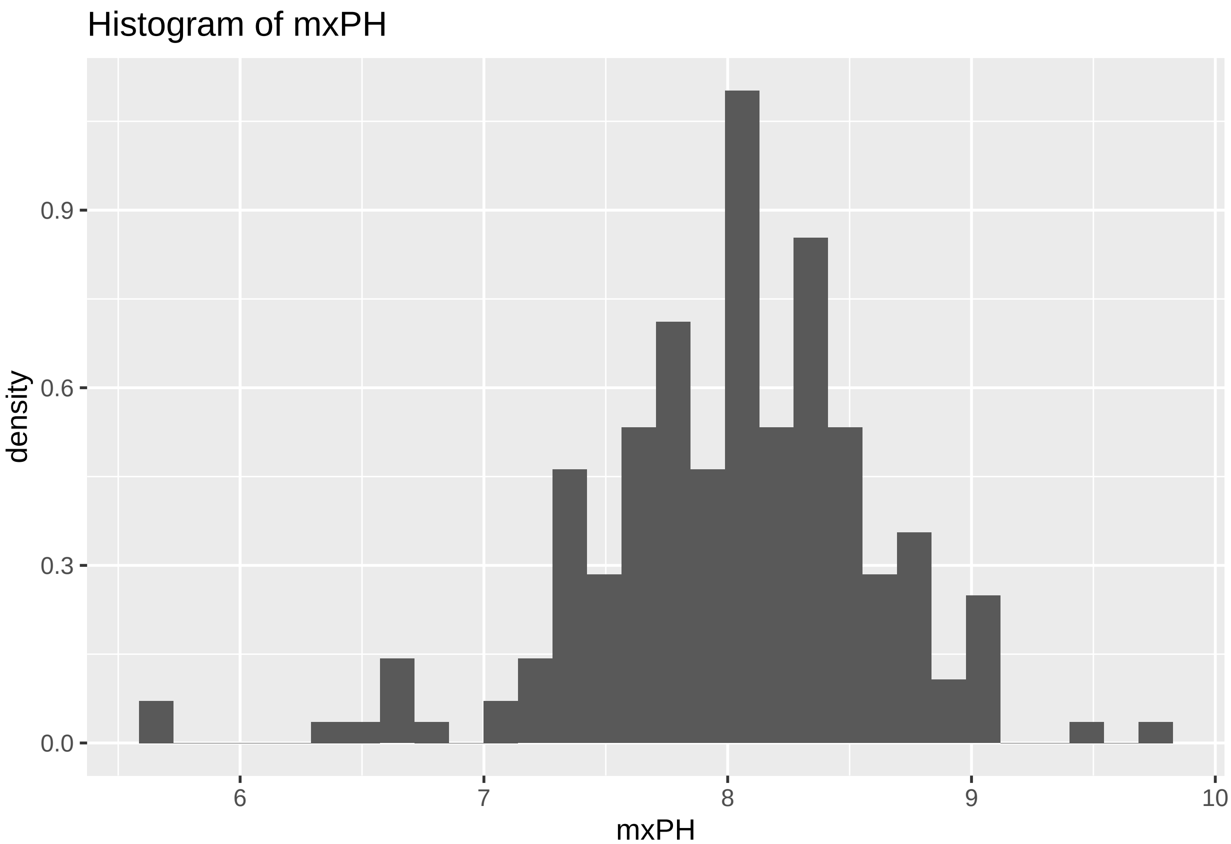
```
## # A tibble: 1 x 8
##    mxPH  mnO2    Cl   NO3   NH4  oPO4   PO4  Chla
##   <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>
## 1 0.504  2.05  33.2  2.17  112.  44.0  122.  6.67
```

The medians are typically smaller than the calculated mean in the question above expect for in the cases of mxPH n=and mnO2 again. The magnitude between the median and MAD is typically much smaller than the magnitude between the mean and variance

2a)

```
mxPH_hist <- algae%>% ggplot(aes(x=mxPH))+geom_histogram(mapping=aes(y=..density..),na.rm = TRUE)+ggtitl
mxPH_hist
```

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

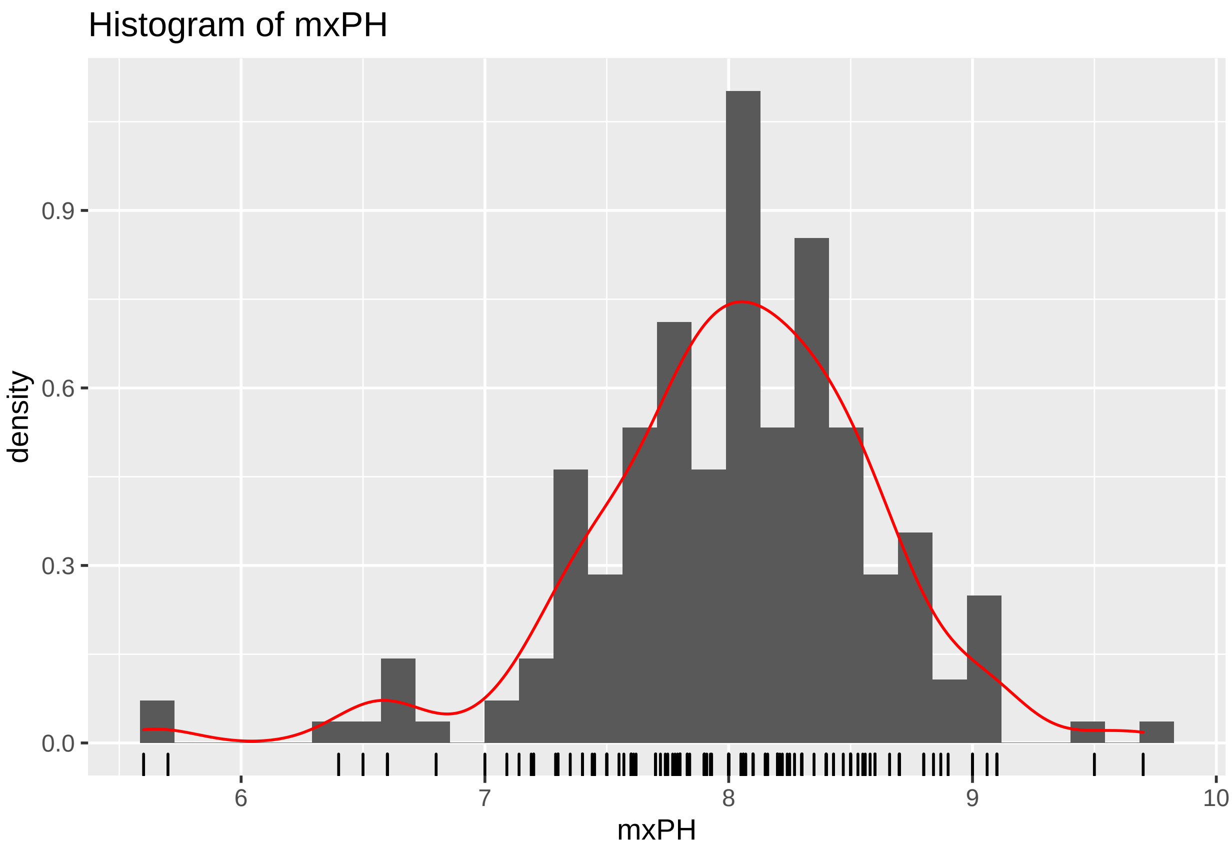


The distribution does not look skewed

2b)

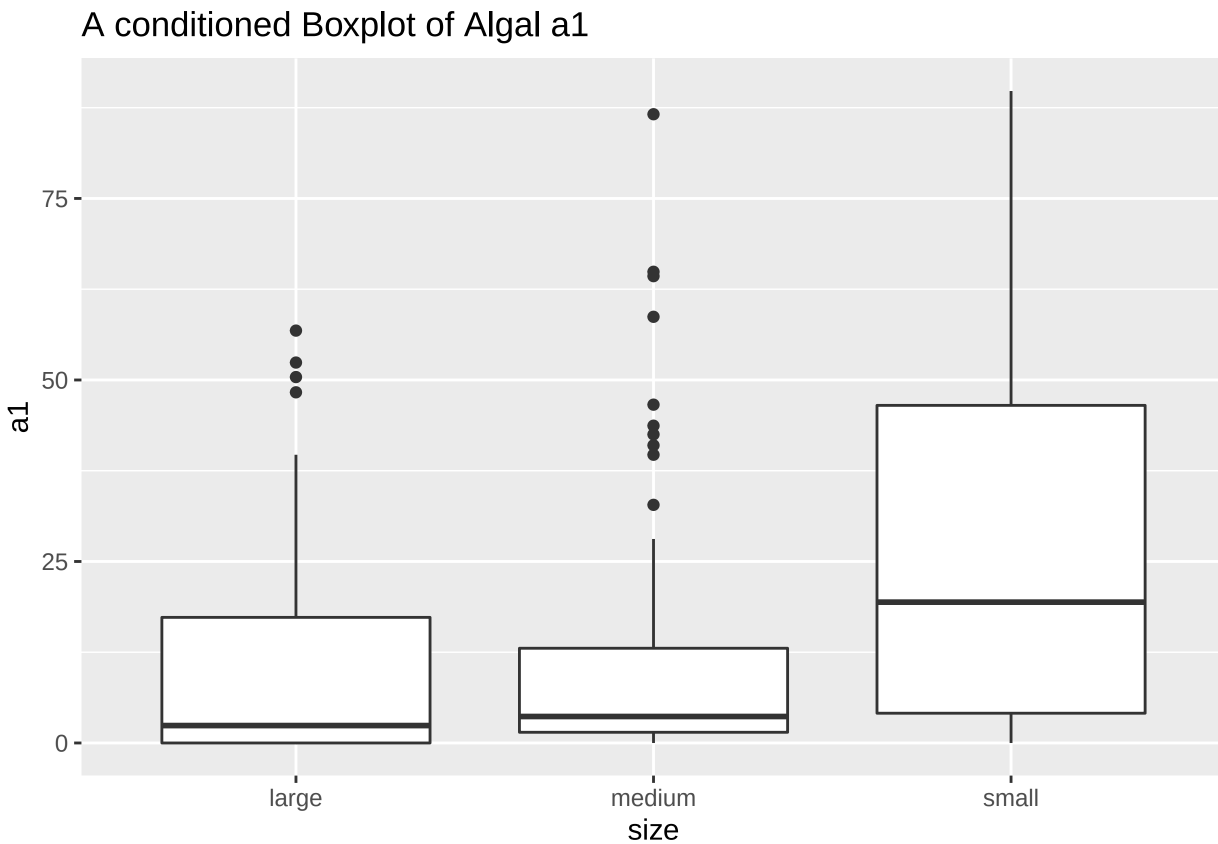
```
mxPH_dens_rug <- mxPH_hist+geom_density(mapping = aes(x=mxPH,y=..density..),na.rm=TRUE,color='red')+geom
mxPH_dens_rug
```

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



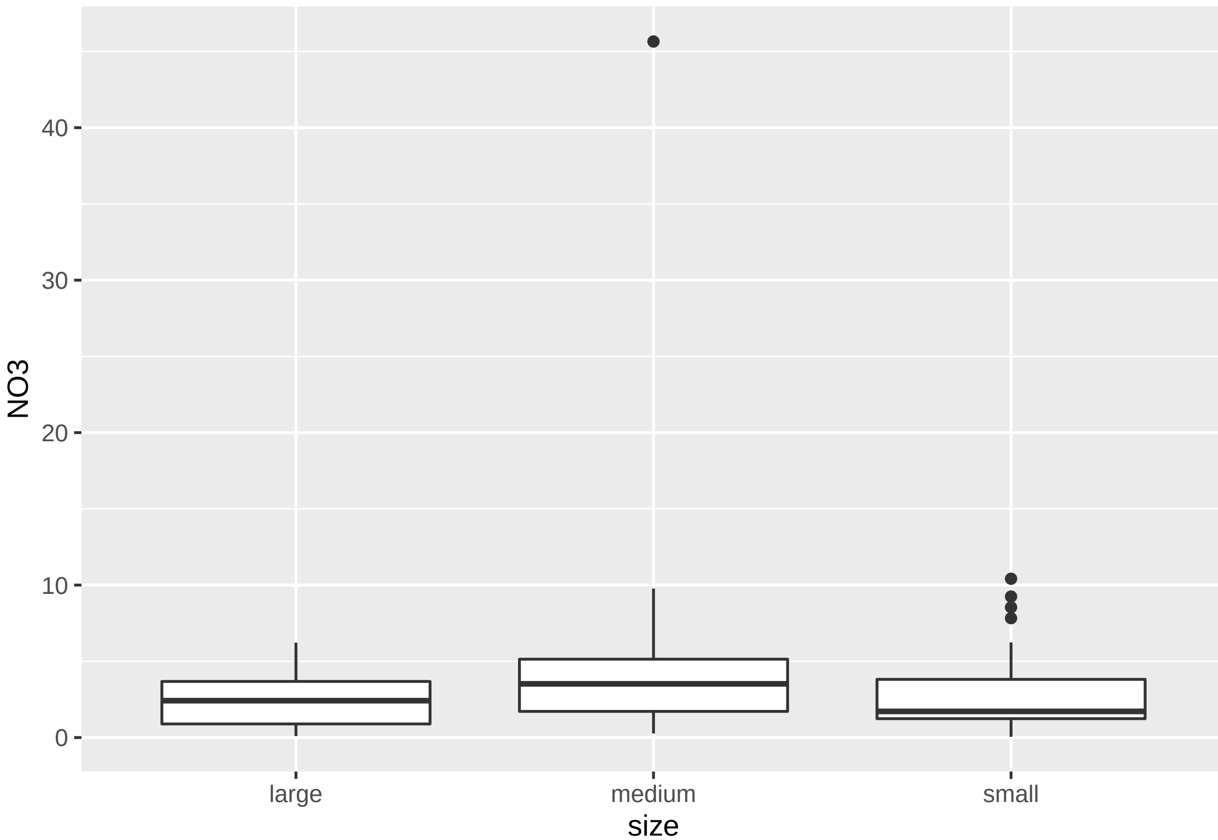
2c)

```
algal_box <- ggplot(data = algae)+geom_boxplot(mapping = aes(x=size,y=a1),na.rm=TRUE)+ggtitle('A conditi
algal_box
```

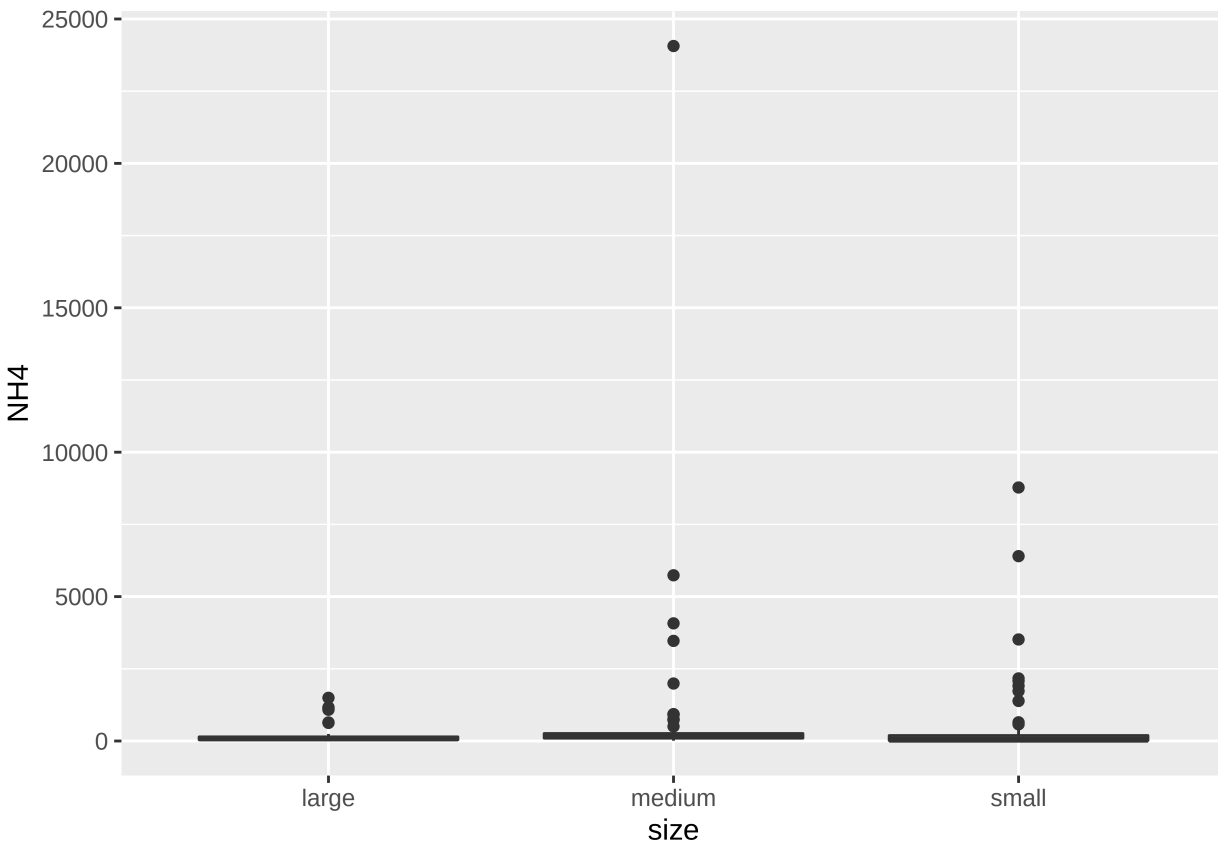


2d)

```
no3 <- ggplot(data = algae)+geom_boxplot(mapping = aes(x=size,y=NO3),na.rm=TRUE)
nh4 <- ggplot(data = algae)+geom_boxplot(mapping = aes(x=size,y=NH4),na.rm=TRUE)
no3
```




```
nh4
```



Using a boxplot, dots that are away from the box depicting mean and median typically stand for outliers. For NO3, there looks to be 1 outlier in the medium size and about 4 in the small size. For NH4, there looks to be about 3 outliers in the large size, 7 in the medium and 7 in the small size.

2e)

```
NO3_Values <- c(mean(algae$NO3,na.rm=TRUE),var(algae$NO3,na.rm=TRUE),median(algae$NO3,na.rm=TRUE),mad(al
NH4_Values <- c(mean(algae$NH4,na.rm=TRUE),var(algae$NH4,na.rm=TRUE),median(algae$NH4,na.rm=TRUE),mad(al
NO3_NH4 <- rbind(NO3_Values,NH4_Values)
colnames(NO3_NH4) <- c('mean','variance','median','MAD')
rownames(NO3_NH4) <- c('NO3','NH4')
NO3_NH4
```

```
##           mean     variance   median        MAD
## NO3   3.282389 1.426176e+01   2.6750   2.172009
## NH4 501.295828 3.851585e+06 103.1665 111.617548
```

The values of NH4 are much greater than those of NO3. The variance is the value most affected when outliers are present as we can see in NH4 which had more outliers as well as an outlier that was very far from the mean

3a)

```
algae %>%
  select(season:a1) %>%
  summarise_all(funs(sum(is.na(.))))
```

```
## # A tibble: 1 x 12
##   season  size speed  mxPH  mnO2    Cl   NO3   NH4  oPO4   PO4  Chla    a1
##    <int> <int> <int> <int> <int> <int> <int> <int> <int> <int> <int> <int>
## 1      0     0     0     1     2    10     2     2     2     2    12     0
```

16 observations and 8 variables contain missing data. mxPH has 1, mnO2 has 2, Cl has 10, NO3 has 2, NH4 has 2, oPO4 has 2, PO4 has 2 and Chla has 12.

3b)

```
algae.del <- algae%>%
  select(season:a1)%>%
  filter(complete.cases(.))
algae.del
```

```
## # A tibble: 184 x 12
##    season size  speed   mxPH  mnO2    Cl    NO3   NH4  oPO4   PO4  Chla    a1
##    <chr>  <chr> <chr>  <dbl> <dbl> <dbl>  <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>
##  1 winter small medium  8      9.8  60.8  6.24  578   105   170   50       0
##  2 spring small medium  8.35   8    57.8  1.29  370   429.  559.   1.3     1.4
##  3 autumn small medium  8.1   11.4  40.0  5.33  347.  126.  187.  15.6     3.3
##  4 spring small medium  8.07   4.8  77.4  2.30   98.2  61.2 139.   1.4     3.1
##  5 autumn small medium  8.06   9    55.4 10.4   234.   58.2  97.6 10.5     9.2
##  6 winter small high    8.25  13.1  65.8  9.25  430    18.2  56.7 28.4    15.1
##  7 summer small high    8.15  10.3  73.2  1.54  110    61.2 112.   3.2     2.4
##  8 autumn small high    8.05  10.6  59.1  4.99  206.   44.7  77.4  6.9    18.2
##  9 winter small medium  8.7    3.4  22.0  0.886 103.   36.3  71    5.54   25.4
## 10 winter small high    7.93   9.9   8    1.39    5.8  27.2  46.6  0.8    17
## # ... with 174 more rows
```

there are now 184 observations which makes sense as we started with 200 and 16 had missing data.

3c)

```
algae.med <- algae%>%
  select(season:a1)%>%
  mutate_at(c('mxPH','mnO2','Cl','NO3','NH4','oPO4','PO4','Chla'),funs(ifelse(is.na(.),median(.,na.rm=T
filter(algae.med,row_number()==48|row_number()==62|row_number()==199)
```

```
## # A tibble: 3 x 12
##   season size  speed   mxPH  mnO2    Cl   NO3   NH4  oPO4   PO4  Chla    a1
##   <chr>  <chr> <chr>  <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>
## 1 winter small low     8.06  12.6   9    0.23   10     5      6   1.1   35.5
## 2 summer small medium  6.4    9.8  32.7  2.68  103.  40.2    14   5.48  19.4
## 3 winter large medium  8      7.6  32.7  2.68  103.  40.2  103.   5.48   0
```

there are 200 observations in algae.med

3d)

```
algae.corr <- algae.del%>%
  select(mxPH:Chla)
cor(algae.corr) #strong correlation with oPO4 and PO4
```

```
##             mxPH        mnO2          Cl        NO3         NH4        oPO4
## mxPH  1.00000000 -0.10269374  0.14709539 -0.1721302 -0.15429757  0.09022909
## mnO2 -0.10269374  1.00000000 -0.26324536  0.1179077 -0.07826816 -0.39375269
## Cl    0.14709539 -0.26324536  1.00000000  0.2109583  0.06598336  0.37925596
## NO3  -0.17213024  0.11790769  0.21095831  1.0000000  0.72467766  0.13301452
## NH4  -0.15429757 -0.07826816  0.06598336  0.7246777  1.00000000  0.21931121
## oPO4  0.09022909 -0.39375269  0.37925596  0.1330145  0.21931121  1.00000000
## PO4   0.10132957 -0.46396073  0.44519118  0.1570297  0.19939575  0.91196460
## Chla  0.43182377 -0.13121671  0.14295776  0.1454929  0.09120406  0.10691478
##             PO4        Chla
```

```
## mxPH  0.1013296  0.43182377
## mnO2 -0.4639607 -0.13121671
## Cl    0.4451912  0.14295776
## NO3   0.1570297  0.14549290
## NH4   0.1993958  0.09120406
## oPO4  0.9119646  0.10691478
## PO4   1.0000000  0.24849223
## Chla  0.2484922  1.00000000
```

```
PO4.oPO4.lm <- lm(algae$PO4~algae$oPO4)
filter(algae,row_number()==28)
```

```
## # A tibble: 1 x 18
##   season size  speed  mxPH  mnO2    Cl   NO3   NH4  oPO4   PO4  Chla    a1    a2
##   <chr>  <chr> <chr> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>
## 1 autumn small high    6.8  11.1     9  0.63    20     4    NA   2.7  30.3   1.9
## # ... with 5 more variables: a3 <dbl>, a4 <dbl>, a5 <dbl>, a6 <dbl>, a7 <dbl>
```

```
#value for oPO4 is 4
predicted.value <- predict(PO4.oPO4.lm,data.frame(p=4),interval='confidence',
                           level = 0.95,type='response')
```

```
## Warning: 'newdata' had 1 row but variables found have 200 rows
```

```
algae$PO4[28] <- predicted.value[28]
algae$PO4[28]
```

```
## [1] 48.06929
```

The predicted value is 48.06929

3e) Using only observed data may lead to wrong assumptions about the actual presense of a chemical. it may be a poor idea to use the median of the chemical for missing values because there may have been an increase or major decrease in the presense of that chemical for that certain observation. It is difficult to know for sure that there is a strong correlation using only the observed data.

4a)

```
obs.ids <- (1:200) #there are 200 observations
chunks <- cut(obs.ids,breaks=5,label=FALSE)%>%sample()
```

4b)

```
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$a1)^2), # compute and store training error
             val.error = mean((predYvl - Yvl$a1)^2)) # compute and store test error
}
```

```r
error <- ldply(1:5,do.chunk,chunkdef=chunks,dat=algae.med)
error
```

```
##   fold train.error val.error
## 1    1    268.3360  376.0809
## 2    2    299.8629  251.1739
## 3    3    276.6203  571.6604
## 4    4    286.8927  324.9421
## 5    5    271.0203  378.5178
```

5a)

```r
algae.Test <- read_table2('algaeTest.txt',col_names=c('season','size','speed','mxPH','mnO2','Cl','NO3',
```

```
##
## -- 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()
## )
```

```r
test.lm <- lm(a1~.,data = algae.Test)
test.error <- mean(((predict(test.lm,algae.Test)-algae.Test$a1)^2))
test.error
```

```
## [1] 218.2218
```

The true test error is 218.2218 which is roughly around the values we estimated in part 4. While it is on the lower side, this was expected because as it is new data, it was not used to train the previous model.

6a)

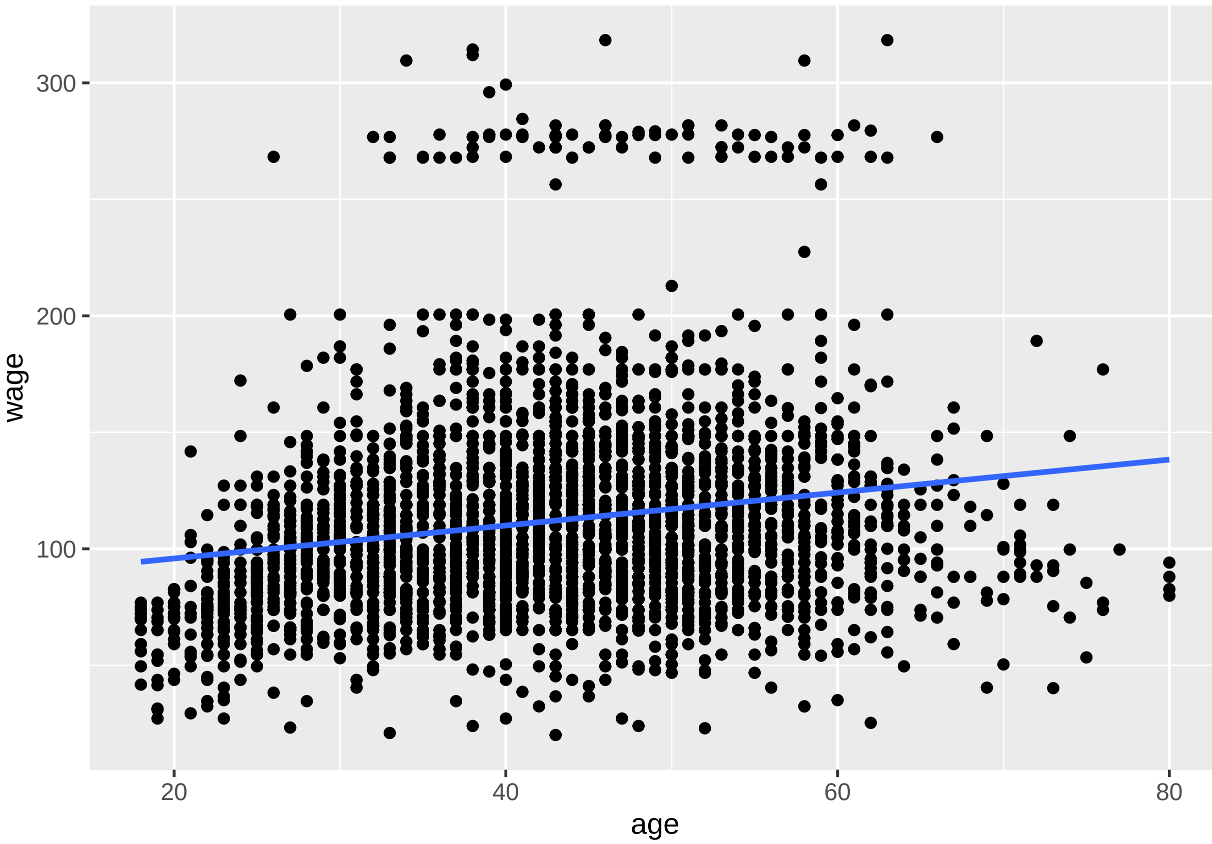
```r
library(ISLR)
head(Wage)
```

```
##        year age            maritl     race       education             region
## 231655 2006  18 1. Never Married 1. White    1. < HS Grad 2. Middle Atlantic
## 86582  2004  24 1. Never Married 1. White 4. College Grad 2. Middle Atlantic
## 161300 2003  45        2. Married 1. White 3. Some College 2. Middle Atlantic
## 155159 2003  43        2. Married 3. Asian 4. College Grad 2. Middle Atlantic
## 11443  2005  50       4. Divorced 1. White      2. HS Grad 2. Middle Atlantic
## 376662 2008  54        2. Married 1. White 4. College Grad 2. Middle Atlantic
##              jobclass         health health_ins  logwage      wage
## 231655  1. Industrial      1. <=Good      2. No 4.318063  75.04315
## 86582  2. Information 2. >=Very Good      2. No 4.255273  70.47602
## 161300  1. Industrial      1. <=Good     1. Yes 4.875061 130.98218
## 155159 2. Information 2. >=Very Good     1. Yes 5.041393 154.68529
## 11443  2. Information      1. <=Good     1. Yes 4.318063  75.04315
```

```
## 376662 2. Information 2. >=Very Good     1. Yes 4.845098 127.11574
```

```r
ggplot(data=Wage,mapping=aes(x=age,y=wage))+geom_point()+geom_smooth(method='lm',se=FALSE)
```

```
## `geom_smooth()` using formula 'y ~ x'
```



From this graph, you can see that wages increase when age increases but they also start to decrease after age 60. This does match what I expect as when one starts working around age 20, they don't start making the most amount of money but as they get older, promotions and raises start coming. Age 60 is around the time that people start to retire so it makes sense that wages goes down after that.

6bi)

```r
wage.lm <- lm(wage~poly(age,degree=10,raw=FALSE),data = Wage)
summary(wage.lm)
```

```
##
## Call:
## lm(formula = wage ~ poly(age, degree = 10, raw = FALSE), data = Wage)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -100.38  -24.45   -4.97   15.49  199.61
##
## Coefficients:
##                                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)                           111.7036     0.7283 153.369  < 2e-16 ***
## poly(age, degree = 10, raw = FALSE)1  447.0679    39.8924  11.207  < 2e-16 ***
## poly(age, degree = 10, raw = FALSE)2 -478.3158    39.8924 -11.990  < 2e-16 ***
## poly(age, degree = 10, raw = FALSE)3  125.5217    39.8924   3.147  0.00167 **
```

```
## poly(age, degree = 10, raw = FALSE)4   -77.9112    39.8924  -1.953  0.05091 .
## poly(age, degree = 10, raw = FALSE)5   -35.8129    39.8924  -0.898  0.36940
## poly(age, degree = 10, raw = FALSE)6    62.7077    39.8924   1.572  0.11607
## poly(age, degree = 10, raw = FALSE)7    50.5498    39.8924   1.267  0.20520
## poly(age, degree = 10, raw = FALSE)8   -11.2547    39.8924  -0.282  0.77787
## poly(age, degree = 10, raw = FALSE)9   -83.6918    39.8924  -2.098  0.03599 *
## poly(age, degree = 10, raw = FALSE)10    1.6240    39.8924   0.041  0.96753
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 39.89 on 2989 degrees of freedom
## Multiple R-squared:  0.08912,    Adjusted R-squared:  0.08607
## F-statistic: 29.24 on 10 and 2989 DF,  p-value: < 2.2e-16
```

6bii)

```
wage.chunk <- cut(1:nrow(Wage),breaks = 5,labels=FALSE)%>%sample()
do.chunk.2 <- function(chunkid,chunkdef,dat,p){
  train = (chunkdef != chunkid)
  Xtr = dat[train,]
  Ytr = dat[train,]
  Xvl = dat[!train,]
  Yvl = dat[!train,]
  if(p==0)
    lm.wage <- lm(wage~1, data = dat[train,])
  else
    lm.wage<- lm(wage~poly(age,degree=p,raw=FALSE),data = dat[train,])
  predYtr = predict(lm.wage)
  predYvl = predict(lm.wage,Xvl)

  data.frame(fold = chunkid,
             train.error = mean((predYtr - Ytr$wage)^2),
             test.error = mean((predYvl - Yvl$wage)^2))
}

error.folds <- NULL
for(j in 0:10){
  tmp <- ldply(1:5,do.chunk.2,chunkdef=wage.chunk,dat=Wage,p=j)
  tmp$degree <- j
  error.folds <- rbind(error.folds,tmp)
}
error.folds
```

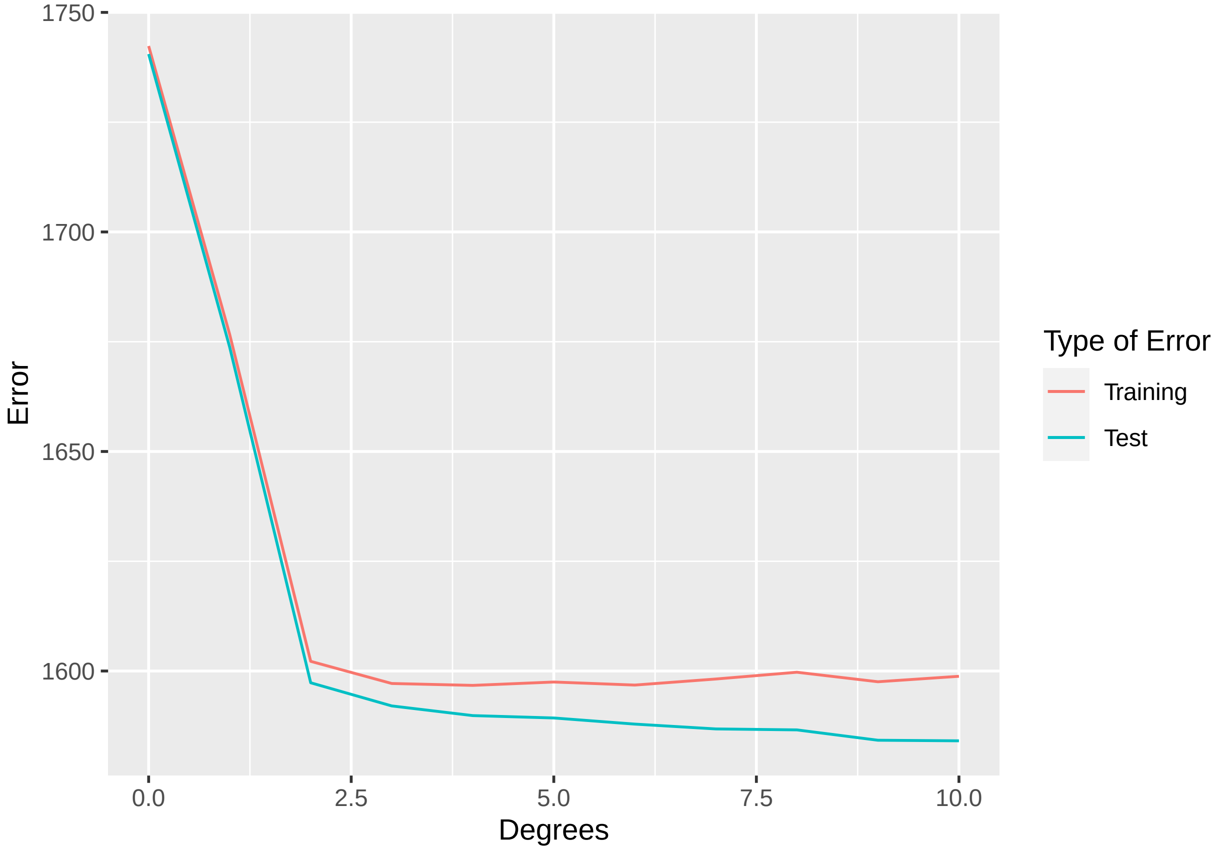
```
##    fold train.error test.error degree
## 1     1    1755.549   1682.551      0
## 2     2    1770.509   1622.570      0
## 3     3    1693.308   1930.530      0
## 4     4    1788.022   1551.485      0
## 5     5    1695.173   1924.584      0
## 6     1    1691.250   1606.696      1
## 7     2    1699.988   1572.095      1
## 8     3    1635.550   1829.689      1
## 9     4    1716.768   1503.703      1
## 10    5    1625.357   1871.232      1
## 11    1    1611.575   1543.897      2
```

```
## 12    2    1610.656   1550.129      2
## 13    3    1556.457   1764.942      2
## 14    4    1653.498   1378.244      2
## 15    5    1554.425   1773.813      2
## 16    1    1606.413   1538.231      3
## 17    2    1604.980   1546.651      3
## 18    3    1551.405   1758.880      3
## 19    4    1647.870   1375.250      3
## 20    5    1549.576   1766.775      3
## 21    1    1605.322   1533.178      4
## 22    2    1603.949   1541.457      4
## 23    3    1549.276   1757.210      4
## 24    4    1643.026   1387.331      4
## 25    5    1547.663   1764.455      4
## 26    1    1605.228   1532.019      5
## 27    2    1603.588   1540.732      5
## 28    3    1548.921   1756.432      5
## 29    4    1641.473   1394.437      5
## 30    5    1547.311   1763.840      5
## 31    1    1604.103   1529.966      6
## 32    2    1602.564   1538.322      6
## 33    3    1547.252   1756.991      6
## 34    4    1638.708   1398.612      6
## 35    5    1546.917   1760.106      6
## 36    1    1604.030   1528.121      7
## 37    2    1601.744   1537.333      7
## 38    3    1546.358   1756.632      7
## 39    4    1635.618   1410.420      7
## 40    5    1546.310   1758.403      7
## 41    1    1603.925   1528.349      8
## 42    2    1601.259   1540.351      8
## 43    3    1546.358   1756.630      8
## 44    4    1635.398   1411.255      8
## 45    5    1546.029   1762.034      8
## 46    1    1601.925   1524.579      9
## 47    2    1597.798   1542.504      9
## 48    3    1545.171   1750.210      9
## 49    4    1632.379   1411.474      9
## 50    5    1543.907   1759.019      9
## 51    1    1601.858   1525.144     10
## 52    2    1597.787   1542.600     10
## 53    3    1545.092   1750.993     10
## 54    4    1632.304   1412.097     10
## 55    5    1543.462   1763.163     10
```

6c)

```
error.group <- error.folds%>%group_by(degree)%>%summarise_at(vars(train.error,test.error),list(name=mea

ggplot()+geom_line(data=error.group,aes(x=degree,y=test.error_name,color='blue'))+geom_line(data=error.g
```



Both errors decrease with an increase in degrees. Around 2 degrees, the graph becomes steady with a slight decrease. Based on this graph we should choose the model with p=10 because that is where both errors are the lowest.