6: Part 1 - Generalized Linear Models

Environmental Data Analytics | John Fay and Luana Lima | Developed by Kateri Salk

Spring 2022

Objectives

1. Answer questions on M5/A5
2. Answer questions on M6 - GLMs
3. Additional comments on videos - t-test
4. Practice more application GLM to real datasets

Set up

```
library(tidyverse)
#install.packages("agricolae")
library(agricolae)

PeterPaul.chem.nutrients <- read.csv("../Data/Processed/NTL-LTER_Lake_Chemistry_Nutrients_PeterPaul_Pro
# Set date to date format
PeterPaul.chem.nutrients$sampledate <- as.Date(PeterPaul.chem.nutrients$sampledate, format = "%Y-%m-%d")

EPAair <- read.csv("../Data/Processed/EPAair_O3_PM25_NC2021_Processed.csv", stringsAsFactors = TRUE)
# Set date to date format
EPAair$Date <- as.Date(EPAair$Date, format = "%Y-%m-%d")


Litter <- read.csv("../Data/Processed/NEON_NIWO_Litter_mass_trap_Processed.csv", stringsAsFactors = TRU
# Set date to date format
Litter$collectDate <- as.Date(Litter$collectDate , format = "%Y-%m-%d")

# Set theme
mytheme <- theme_classic(base_size = 14) +
  theme(axis.text = element_text(color = "black"),
        legend.position = "top")
theme_set(mytheme)
```

T-Test

Continuous response, one categorical explanatory variable with two categories (or comparison to a single value if a one-sample test).

Formulating Hypothesis for µ

Two hypotheses are formed – the null hypothesis and the alternative hypothesis. The null hypothesis and the alternative hypothesis combine to cover all possible values for the population mean. The null hypothesis must have the equality. The null and alternative hypotheses are always stated in terms of the population mean (mu).

One-sample t-test

The object of a one sample test is to test the null hypothesis that the mean of the group is equal to a specific value. For example, we might ask ourselves (from the EPA air quality processed dataset):

Function t.test() **x** a (non-empty) numeric vector of data values. **alternative** a character string specifying the alternative hypothesis, must be one of "two.sided" (default), "greater" or "less". You can specify just the initial letter. **mu** a number indicating the true value of the mean (or difference in means if you are performing a two sample test). **formula** a formula of the form lhs ~ rhs where lhs is a numeric variable giving the data values and rhs either 1 for a one-sample or paired test or a factor with two levels giving the corresponding groups. If lhs is of class "Pair" and rhs is 1, a paired test is done.

Are Ozone levels below the threshold for "good" AQI index (0-50)?

> Exercise 1: State the hypotheses for testing mean of AQI index.
>
> Answer: Is ozone less than 50 ppm?

```
summary(EPAair$Ozone)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##    5.00   32.00   40.00   40.88   46.00  129.00    2146
```

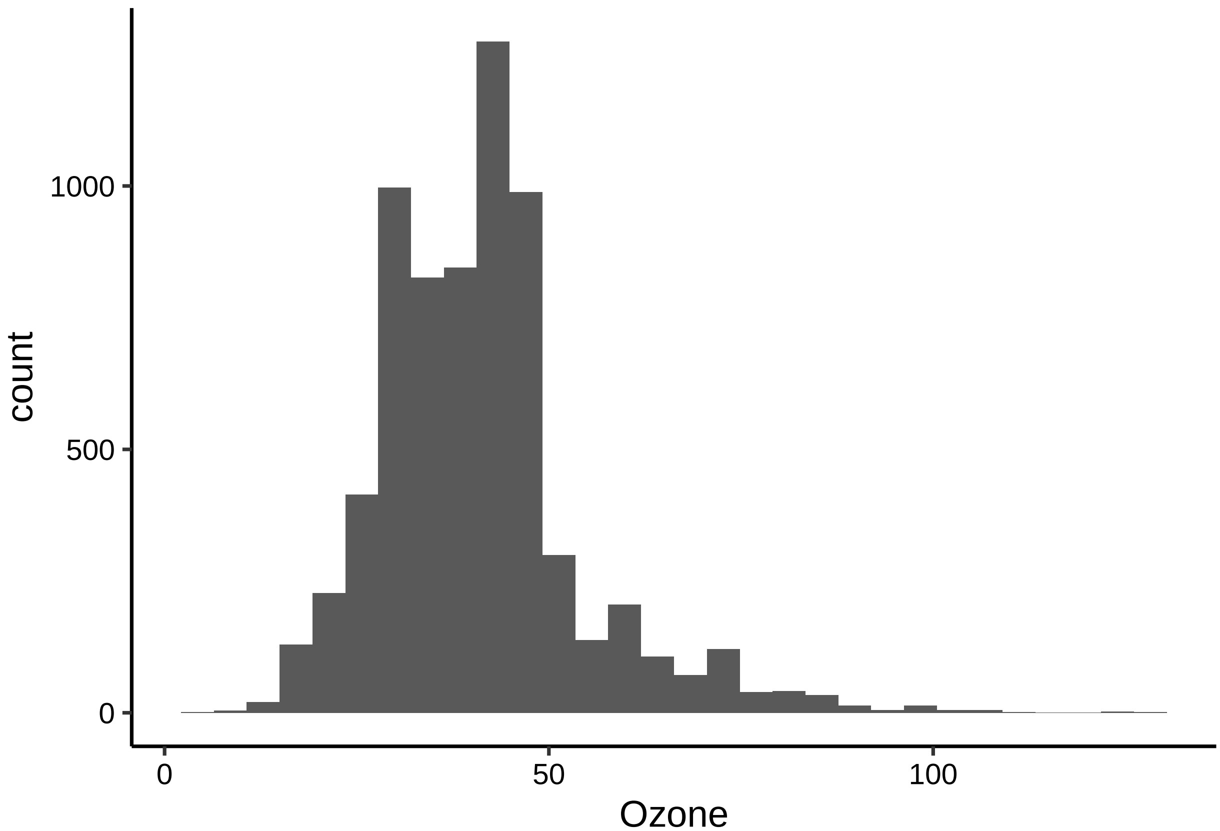
```
EPAair.subsample <- sample_n(EPAair, 5000)

# Evaluate assumption of normal distribution
shapiro.test((EPAair.subsample$Ozone))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  (EPAair.subsample$Ozone)
## W = 0.92044, p-value < 2.2e-16
```

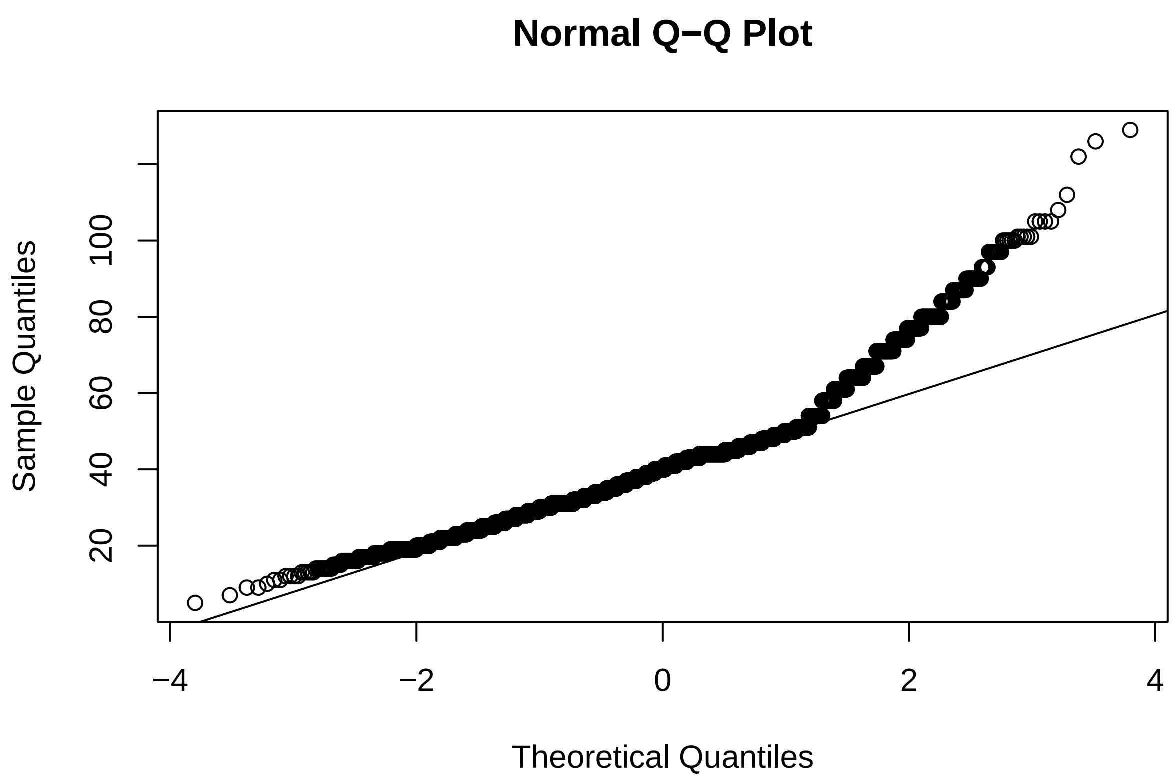
```
ggplot(EPAair, aes(x = Ozone)) +
  geom_histogram()
```

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

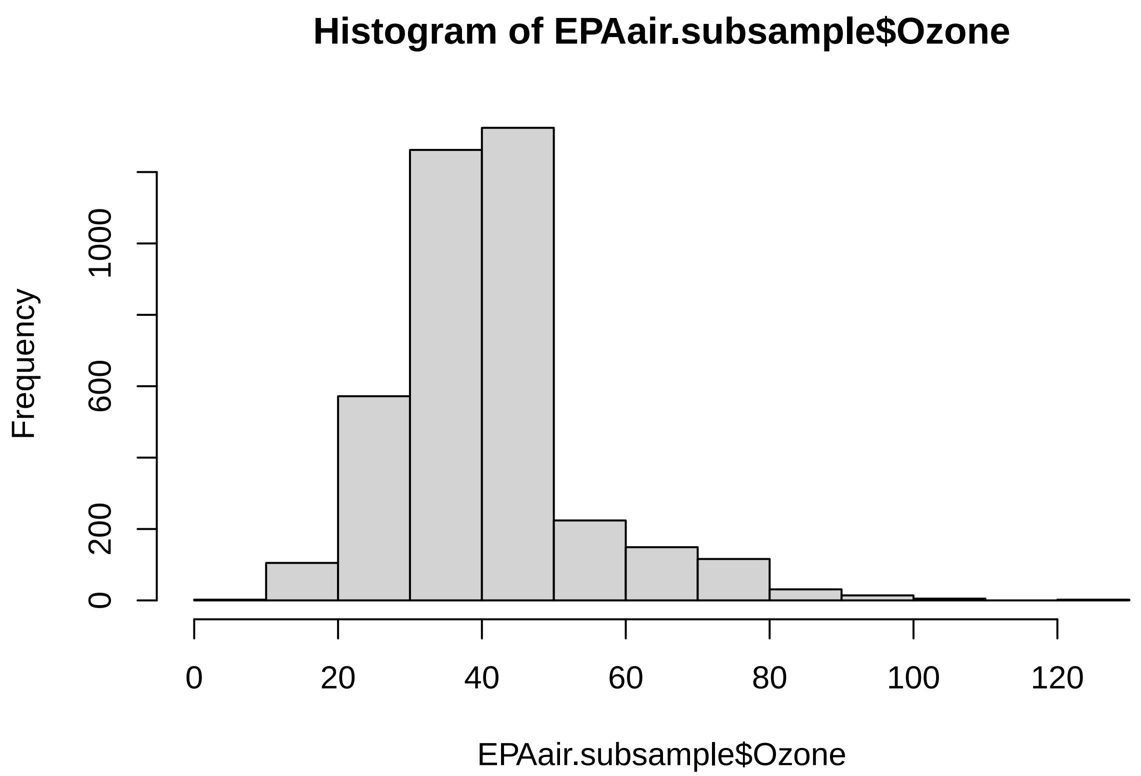
```
## Warning: Removed 2146 rows containing non-finite values (stat_bin).
```




```
qqnorm(EPAair$Ozone); qqline(EPAair$Ozone)
```



```r
#histogram
hist(EPAair.subsample$Ozone)
```

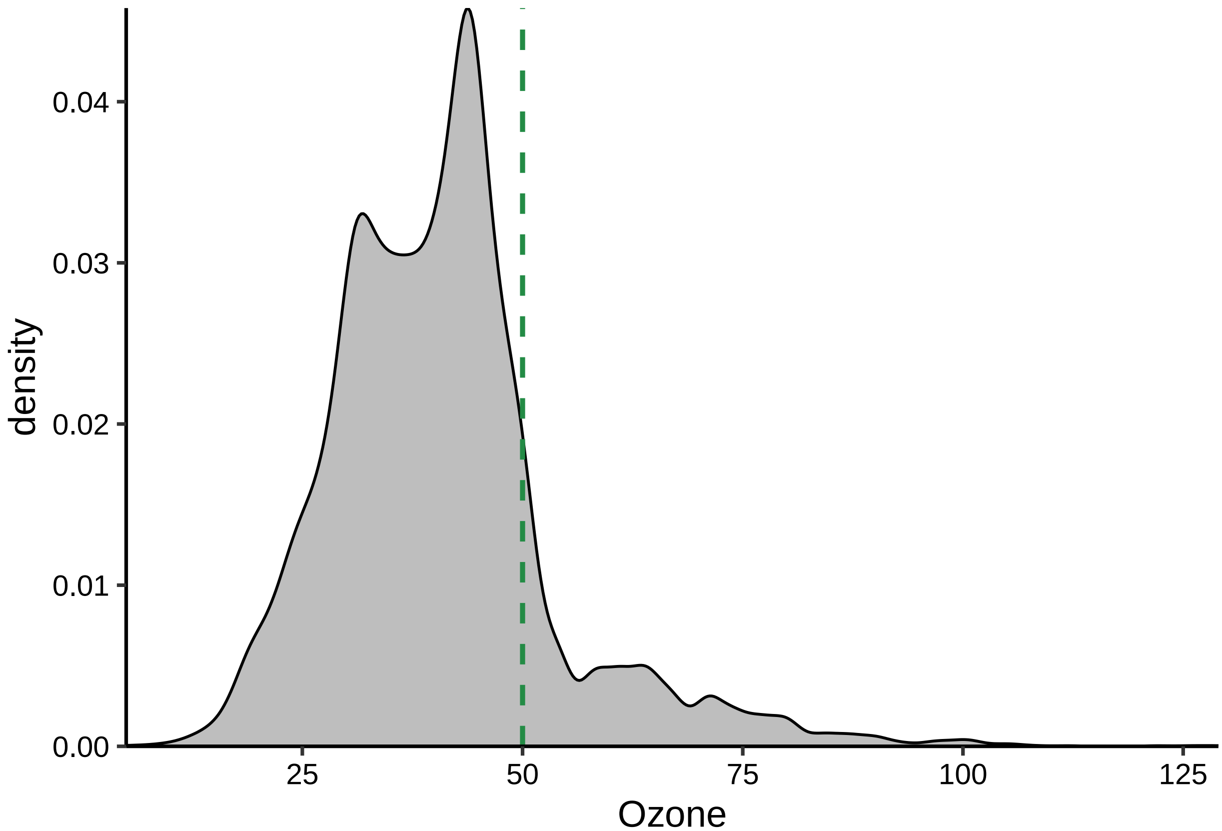



```r
O3.onesample <- t.test(EPAair$Ozone, mu = 50, alternative = "less")
O3.onesample
```

```
##
##  One Sample t-test
##
## data:  EPAair$Ozone
## t = -57.98, df = 6829, p-value < 2.2e-16
## alternative hypothesis: true mean is less than 50
## 95 percent confidence interval:
##      -Inf 41.13416
## sample estimates:
## mean of x
##  40.87526
```

```r
#we are sufficiently below air quality

Ozone.plot <- ggplot(EPAair, aes(x = Ozone)) +
  #geom_density(stat = "count", fill = "gray") +
  geom_density(fill = "gray") +
  geom_vline(xintercept = 50, color = "#238b45", lty = 2, size = 0.9) +
  scale_x_continuous(expand = c(0, 0)) + scale_y_continuous(expand = c(0, 0))
print(Ozone.plot)
```

```
## Warning: Removed 2146 rows containing non-finite values (stat_density).
```



Write a sentence or two about the results of this test. Include both the results of the test and an interpretation that puts the findings in context of the research question.

Two-sample t-test

The two-sample *t* test is used to test the hypothesis that the mean of two samples is equivalent. Unlike the one-sample tests, a two-sample test requires a second assumption that the variance of the two groups is equivalent. Are Ozone levels different between 2018 and 2019?

```
shapiro.test(EPAair$Ozone[EPAair$Year == 2018])
```

```
##
##  Shapiro-Wilk normality test
##
## data:  EPAair$Ozone[EPAair$Year == 2018]
## W = 0.92665, p-value < 2.2e-16
```

```
shapiro.test(EPAair$Ozone[EPAair$Year == 2019])
```

```
##
##  Shapiro-Wilk normality test
##
## data:  EPAair$Ozone[EPAair$Year == 2019]
## W = 0.92132, p-value < 2.2e-16
```

```
#p-value less than 0.05 then reject null for 2018 and 2019 i.e. data do not follow normal distribution

#Compare variance using F-test (only)
```

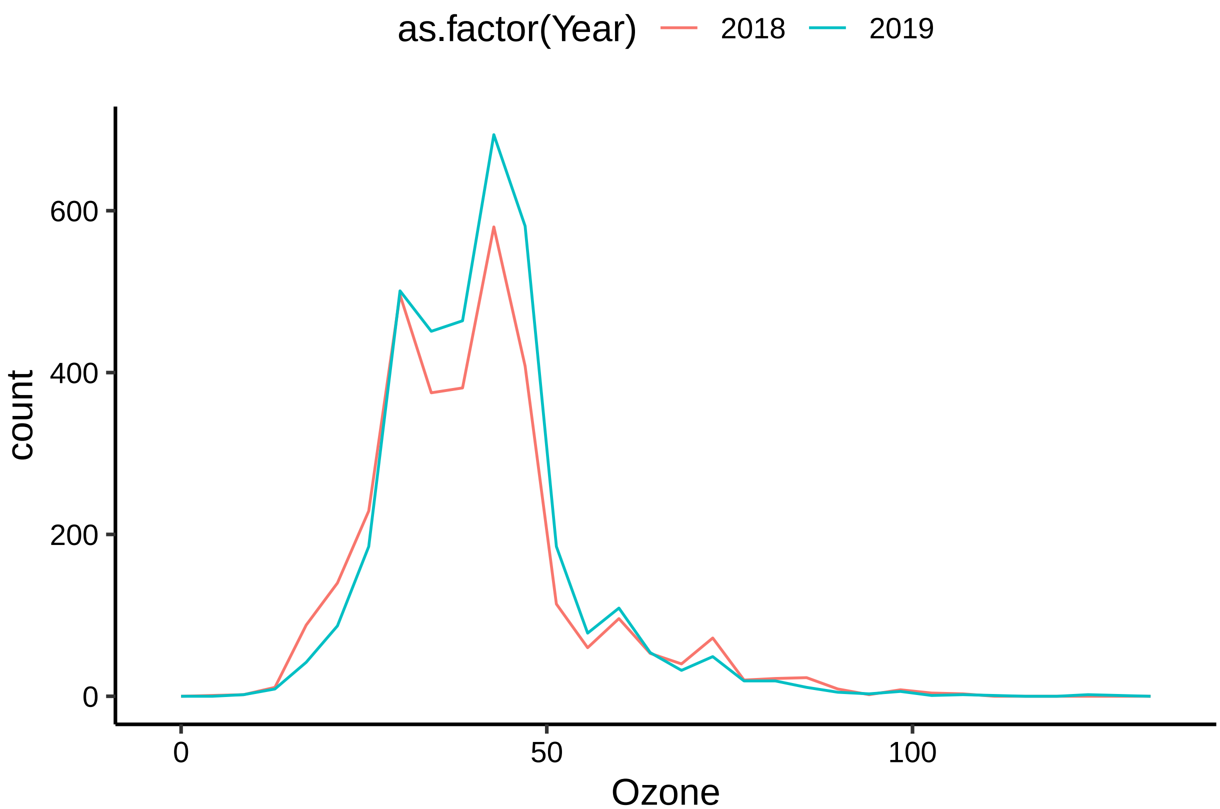
```r
var.test(EPAair$Ozone ~ EPAair$Year)
```

```
##
##  F test to compare two variances
##
## data:  EPAair$Ozone by EPAair$Year
## F = 1.3061, num df = 3236, denom df = 3592, p-value = 6.217e-15
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  1.221265 1.396919
## sample estimates:
## ratio of variances
##           1.306065
```

```r
#p-value less than 0.05 then reject null for 2018 and 2019 i.e. true ratio not equal to one

ggplot(EPAair, aes(x = Ozone, color = as.factor(Year))) +
  geom_freqpoly()
```

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

```
## Warning: Removed 2146 rows containing non-finite values (stat_bin).
```




```r
# Format as a t-test
O3.twosample <- t.test(EPAair$Ozone ~ EPAair$Year)
O3.twosample
```

```
##
##  Welch Two Sample t-test
##
```

```
## data:  EPAair$Ozone by EPAair$Year
## t = -2.6642, df = 6467.7, p-value = 0.007736
## alternative hypothesis: true difference in means between group 2018 and group 2019 is not equal to 0
## 95 percent confidence interval:
##  -1.4670426 -0.2232942
## sample estimates:
## mean in group 2018 mean in group 2019
##           40.43065           41.27581
```

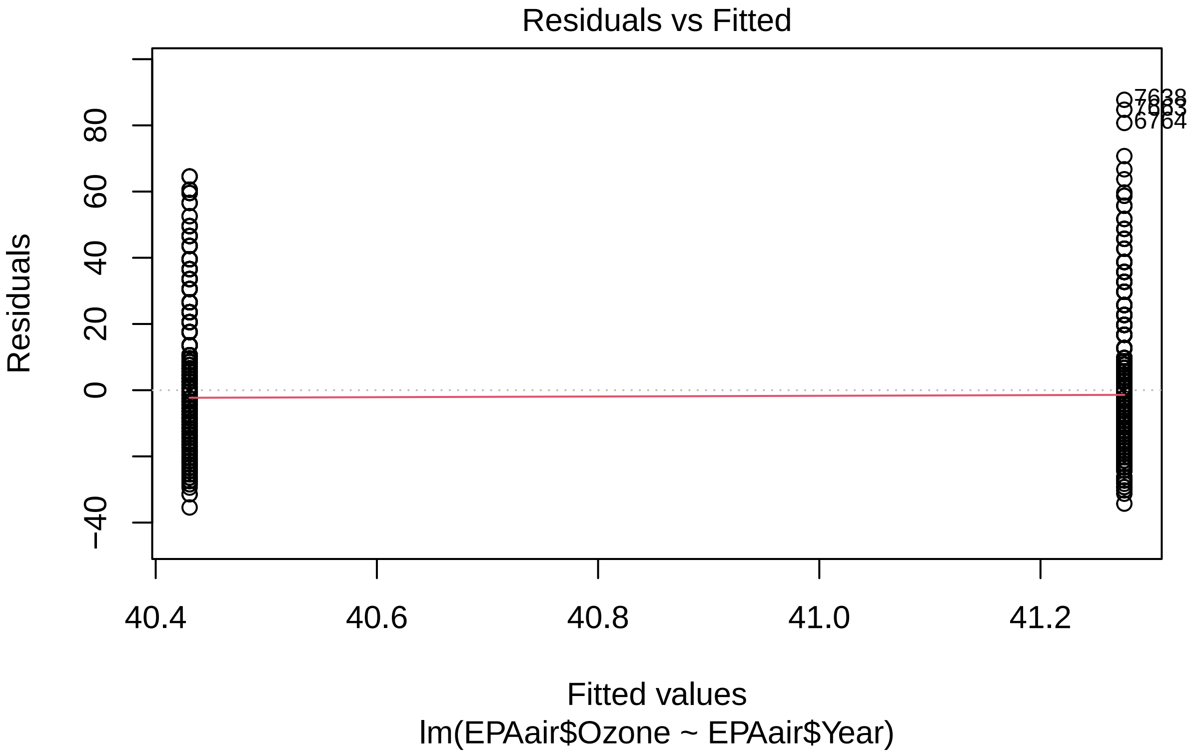
```
O3.twosample$p.value
```

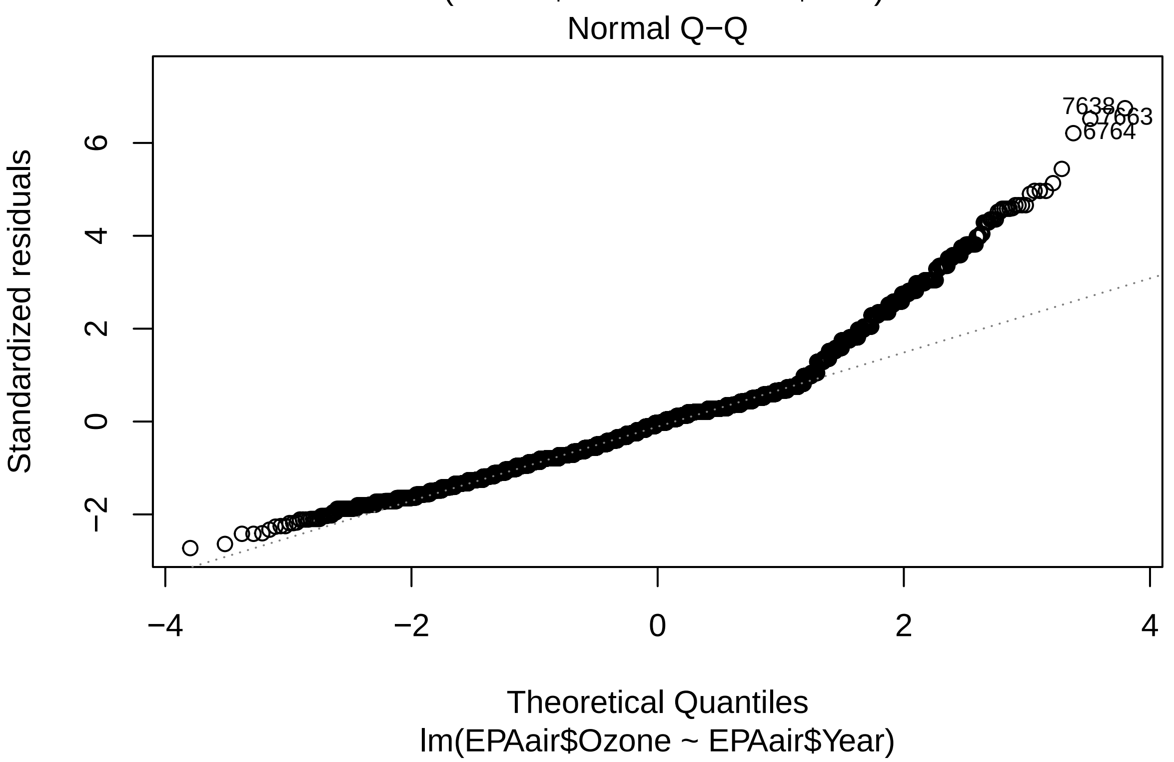
```
## [1] 0.00773585
```

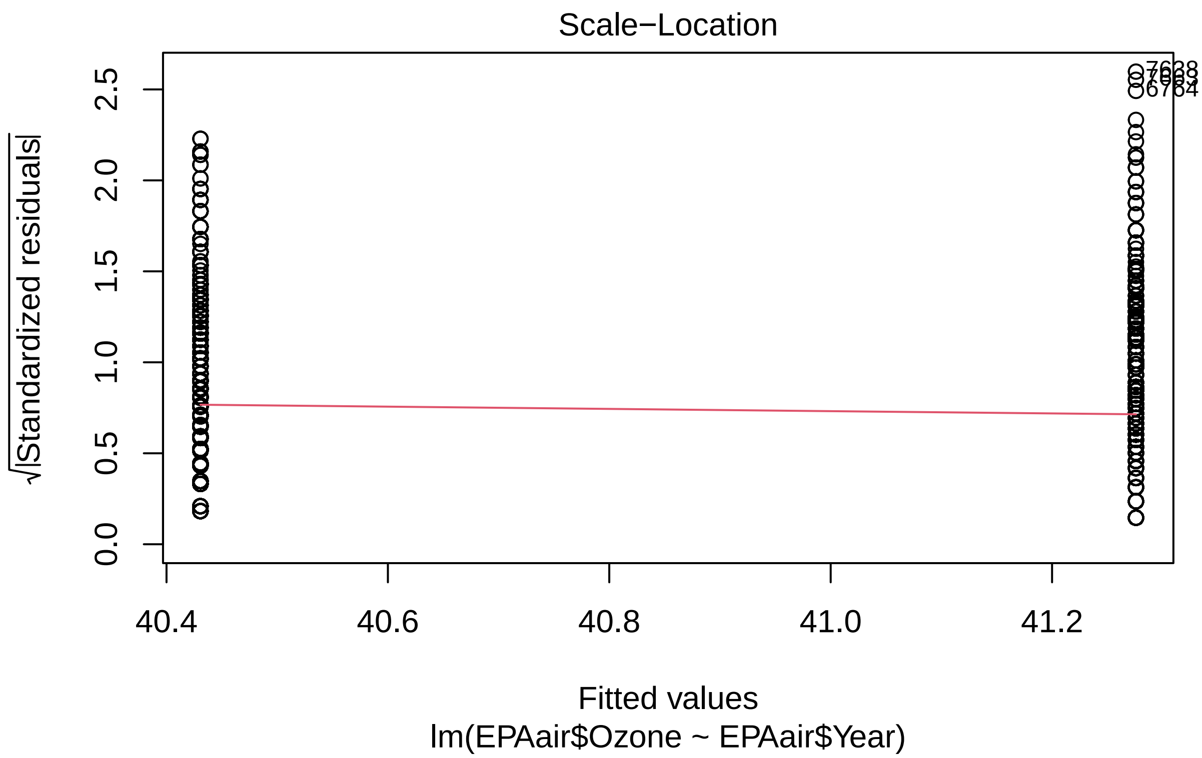
```
# Format as a GLM
O3.twosample2 <- lm(EPAair$Ozone ~ EPAair$Year)
summary(O3.twosample2)
```

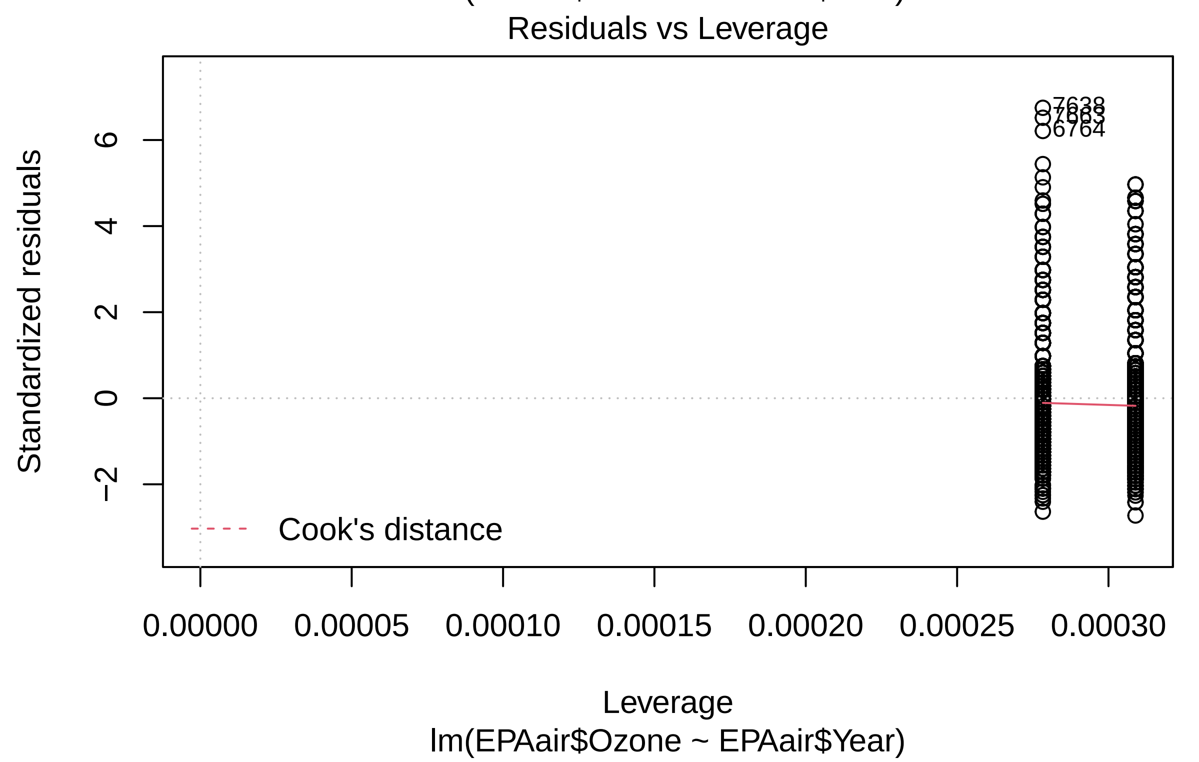
```
##
## Call:
## lm(formula = EPAair$Ozone ~ EPAair$Year)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -35.431  -8.431  -0.431   5.569  87.724
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1665.1192   635.9203  -2.618  0.00885 **
## EPAair$Year     0.8452     0.3150   2.683  0.00732 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 13 on 6828 degrees of freedom
##   (2146 observations deleted due to missingness)
## Multiple R-squared:  0.001053,   Adjusted R-squared:  0.0009066
## F-statistic: 7.197 on 1 and 6828 DF,  p-value: 0.00732
```

```
plot(O3.twosample2)
```


Residuals vs Fitted
Residuals
−40
0
20
40
60
80
40.4
40.6
40.8
41.0
41.2
7638
7663
6764
Fitted values
lm(EPAair$Ozone ~ EPAair$Year)


Normal Q−Q
Standardized residuals
−2
0
2
4
6
−4
−2
0
2
4
7638
7663
6764
Theoretical Quantiles
lm(EPAair$Ozone ~ EPAair$Year)





Statistical Test: Cheat sheet

F-test: Compare the variances of two groups. The data must be normally distributed.

Bartlett's test: Compare the variances of two or more groups. The data must be normally distributed.

Shapiro.test: check for normality

One-sample t-test: check if mean is equal/less/greater to specific value, single variable

Two-sample t-test: check if mean of two samples is equivalent

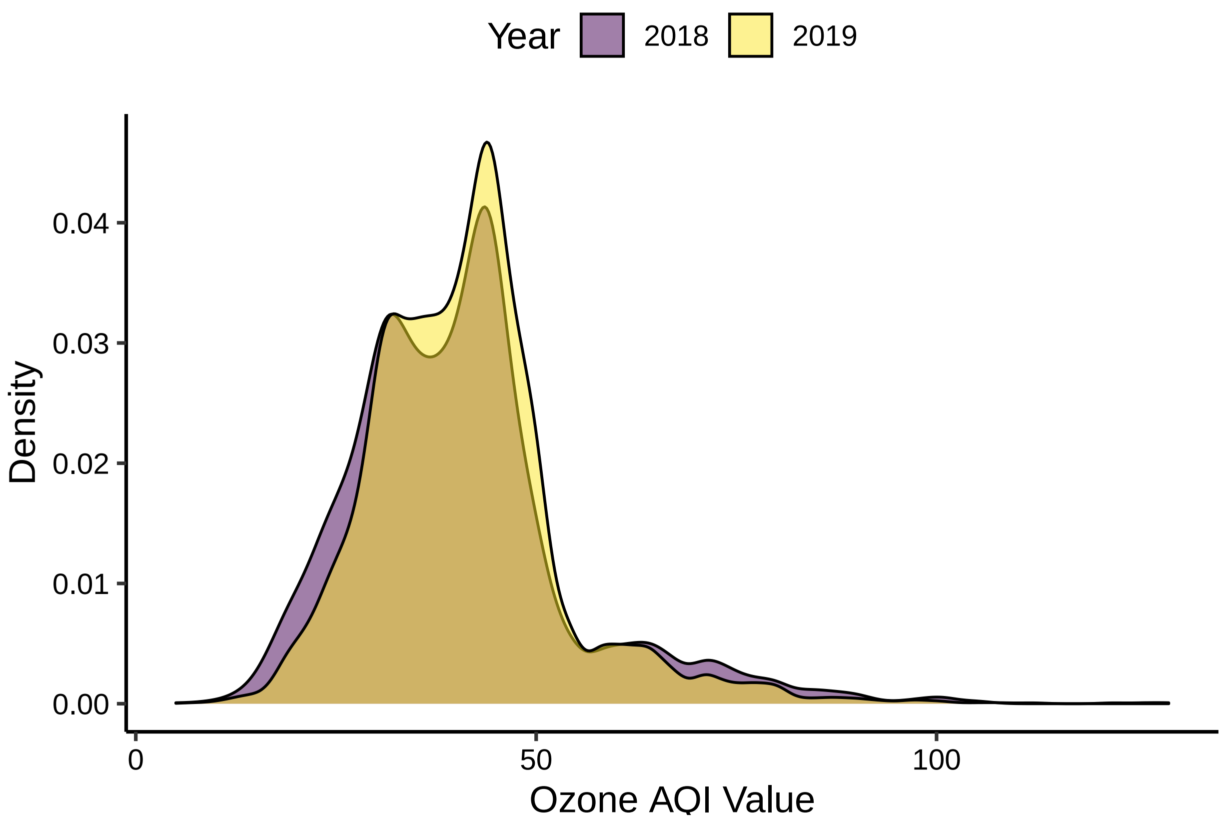
Visualization and interpretation challenge

Create three plots, each with appropriately formatted axes and legends. Choose a non-default color palette.

1. geom_density of ozone divided by year (distinguish between years by adding transparency to the geom_density layer).
2. geom_boxplot of ozone divided by year . Add letters representing a significant difference between 2018 and 2019 (hint: stat_summary).
3. geom_violin of ozone divided by year, with the 0.5 quantile marked as a horizontal line. Add letters representing a significant difference between 2018 and 2019.

```
#Exercise 2:

ggplot(EPAair, aes(x = Ozone, fill = as.factor(Year))) +
  geom_density(alpha = 0.5) +
    scale_fill_viridis_d() +
  labs(x = "Ozone AQI Value", y = "Density", fill = "Year")
```

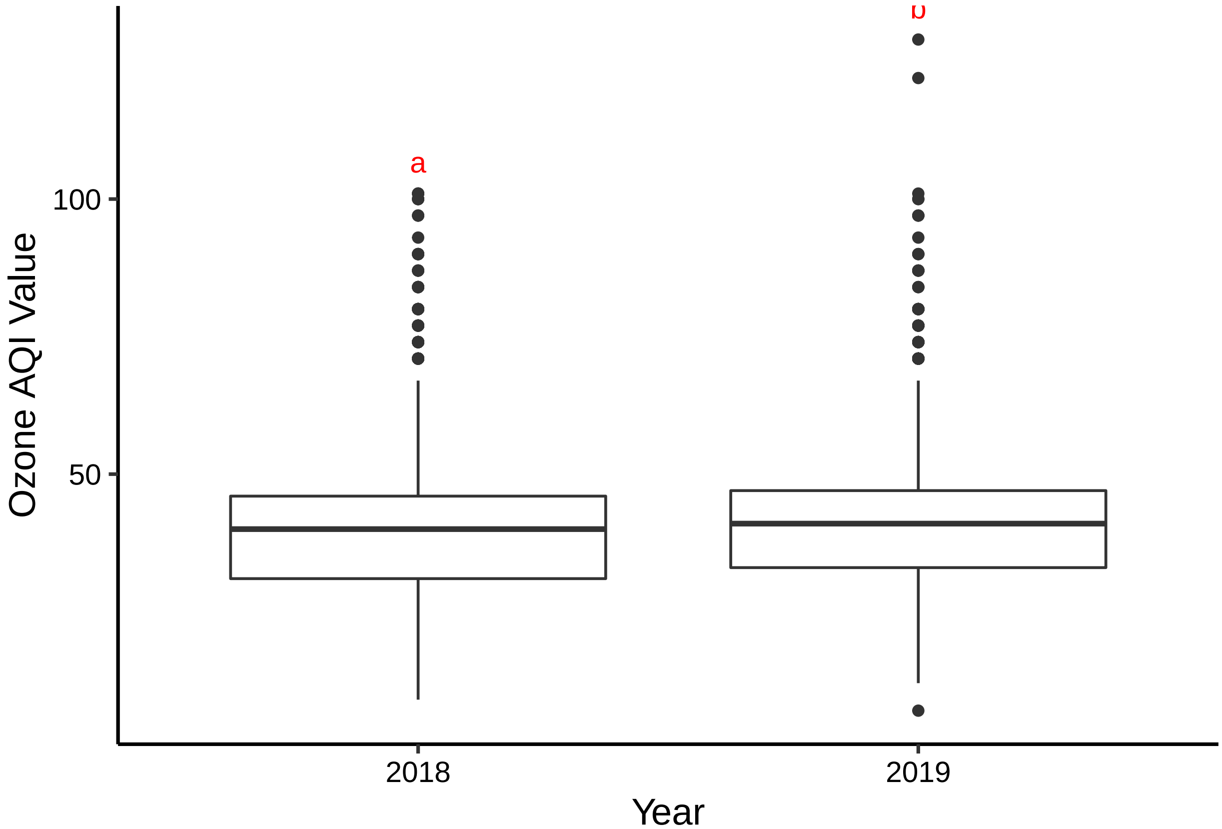
```
## Warning: Removed 2146 rows containing non-finite values (stat_density).
```




```
ggplot(EPAair.subsample, aes(x = as.factor(Year), y = Ozone))+
  geom_boxplot() +
  stat_summary(geom = "text", fun.y = max, vjust = -1, size = 4,
               label = c("a", "b"), color = "red") +
  labs(x = "Year", y = "Ozone AQI Value")
```

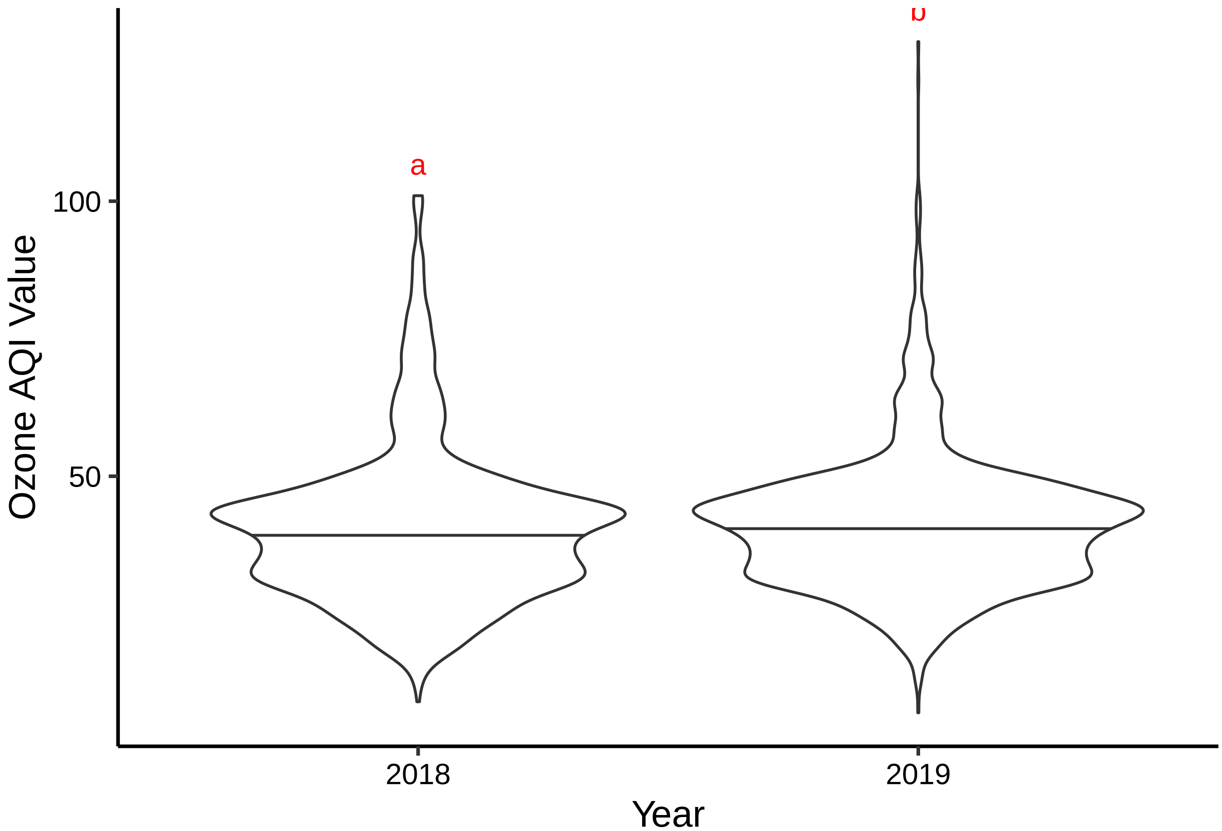
```
## Warning: `fun.y` is deprecated. Use `fun` instead.
```

```
## Warning: Removed 1194 rows containing non-finite values (stat_boxplot).

## Warning: Removed 1194 rows containing non-finite values (stat_summary).
```




```
ggplot(EPAair.subsample, aes(x = as.factor(Year), y = Ozone))+
  geom_violin(draw_quantiles = 0.50) +
  stat_summary(geom = "text", fun = max, vjust = -1, size = 4, color = "red",
               label = c("a", "b")) +
  labs(x = "Year", y = "Ozone AQI Value")
```

```
## Warning: Removed 1194 rows containing non-finite values (stat_ydensity).

## Warning: Removed 1194 rows containing non-finite values (stat_summary).
```



Linear Regression

Important components of the linear regression are the correlation and the R-squared value. The **correlation** is a number between -1 and 1, describing the relationship between the variables. Correlations close to -1 represent strong negative correlations, correlations close to zero represent weak correlations, and correlations close to 1 represent strong positive correlations. The **R-squared value** is the correlation squared, becoming a number between 0 and 1. The R-squared value describes the percent of variance accounted for by the explanatory variables.

For the NTL-LTER dataset, can we predict PM2.5 from Ozone?

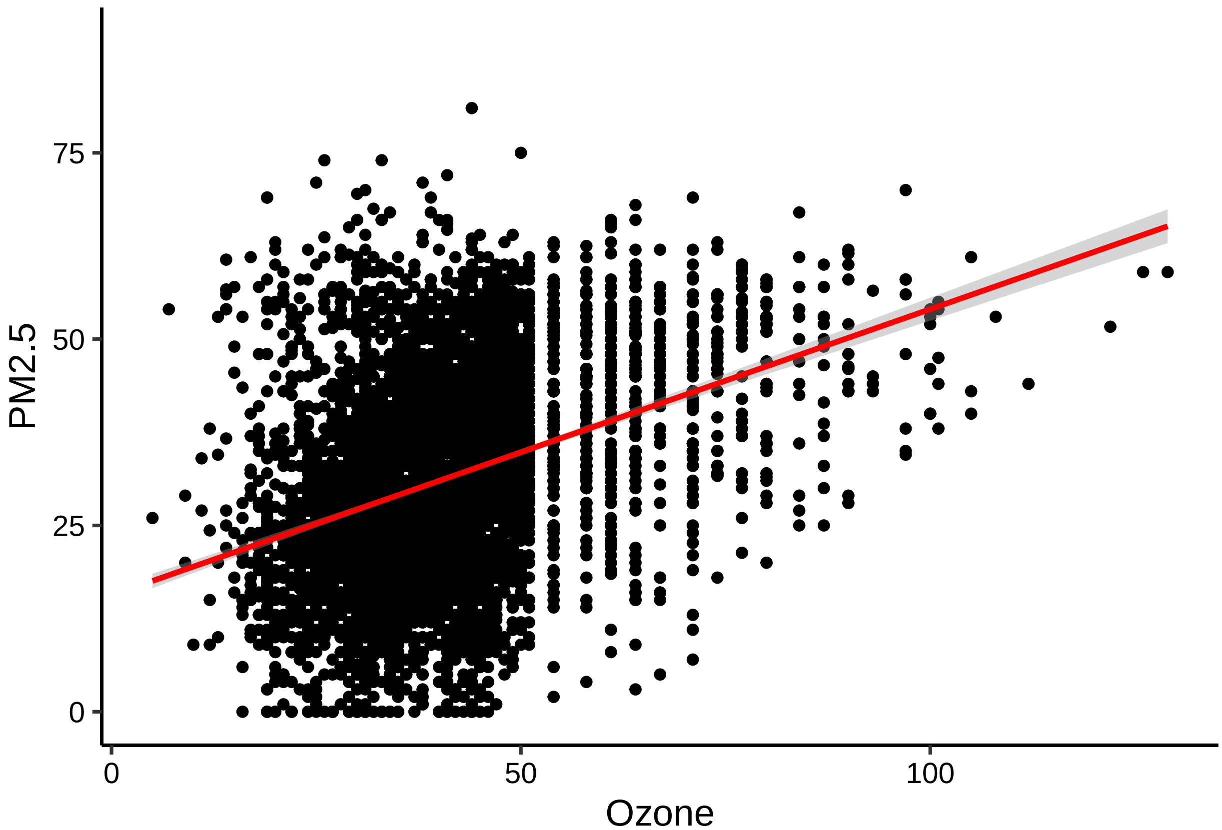
```
#Exercise 3: Run a linear regression PM2.5 by Ozone. Find the p-value and R-squared value.
pm25.ozone_lm <- lm(data = EPAair, PM2.5 ~ Ozone)
summary(pm25.ozone_lm)
```

```
##
## Call:
## lm(formula = PM2.5 ~ Ozone, data = EPAair)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -37.204  -8.931  -0.613   8.463  48.473
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 15.63824    0.55556   28.15   <2e-16 ***
## Ozone        0.38384    0.01298   29.58   <2e-16 ***
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 13.06 on 5774 degrees of freedom
##   (3200 observations deleted due to missingness)
## Multiple R-squared:  0.1316, Adjusted R-squared:  0.1314
## F-statistic: 874.9 on 1 and 5774 DF,  p-value: < 2.2e-16
```

```
# p-value < 2e-16
# r^2 = 0.1314 --> %variance in the dependent variable

#Exercise 4: Build a scatterplot. Add a line and standard error for the linear regression.
ggplot(EPAair, aes(x = Ozone, y = PM2.5)) +
  geom_point() +
  geom_smooth(method = 'lm', se = TRUE, color = "red")
```

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

## Warning: Removed 3200 rows containing non-finite values (stat_smooth).

## Warning: Removed 3200 rows containing missing values (geom_point).
```



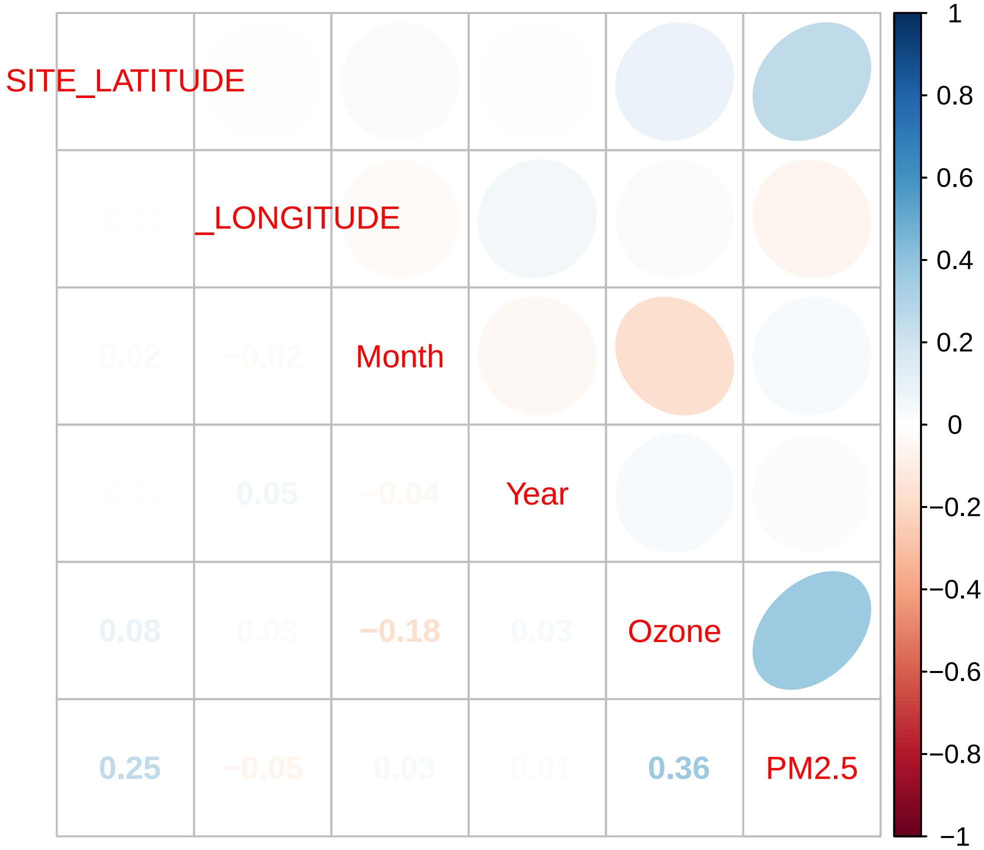
AIC to select variables

What other variables can we add to improve model?

```
#Exercise 5: Build correlation plots and identify more possible explanatory variables to add to the reg
library(corrplot)
```

```
## corrplot 0.92 loaded
```

```
AIC.corrplot <- cor(EPAair %>%
                      select(SITE_LATITUDE:PM2.5) %>%
                      na.omit())
corrplot.mixed(AIC.corrplot, upper = "ellipse")
```




```
#Exercise 6: Choose a model by AIC in a Stepwise Algorithm. Do the results from AIC match the variables

model <- lm(data = EPAair, PM2.5 ~ Ozone + Year + Month + SITE_LATITUDE + SITE_LONGITUDE)
step(model)
```

```
## Start:  AIC=29272.11
## PM2.5 ~ Ozone + Year + Month + SITE_LATITUDE + SITE_LONGITUDE
##
##                  Df Sum of Sq     RSS   AIC
## - Year            1       149  915695 29271
## <none>                         915545 29272
## - SITE_LONGITUDE  1      4087  919632 29296
## - Month           1      8874  924420 29326
## - SITE_LATITUDE   1     54272  969818 29603
## - Ozone           1    142142 1057688 30104
##
## Step:  AIC=29271.05
## PM2.5 ~ Ozone + Month + SITE_LATITUDE + SITE_LONGITUDE
##
##                  Df Sum of Sq    RSS   AIC
## <none>                        915695 29271
## - SITE_LONGITUDE  1      4017 919712 29294
```

```
## - Month           1      8815  924510 29324
## - SITE_LATITUDE   1     54223  969918 29601
## - Ozone           1    142470 1058165 30104
##
## Call:
## lm(formula = PM2.5 ~ Ozone + Month + SITE_LATITUDE + SITE_LONGITUDE,
##     data = EPAair)
##
## Coefficients:
##    (Intercept)           Ozone           Month   SITE_LATITUDE  SITE_LONGITUDE
##      -259.2766          0.3826          0.4643          6.5210         -0.4956
```

```r
#Exercise 7: Run another regression using the variables selected on Exercise 6. Compare r-squared value

final_model <- lm(data = EPAair, PM2.5 ~ Ozone + Month + SITE_LATITUDE + SITE_LONGITUDE)
summary(final_model)
```

```
##
## Call:
## lm(formula = PM2.5 ~ Ozone + Month + SITE_LATITUDE + SITE_LONGITUDE,
##     data = EPAair)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -35.806  -8.846  -0.948   7.777  52.098
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)    -259.27663   14.74368 -17.586  < 2e-16 ***
## Ozone             0.38257    0.01277  29.965  < 2e-16 ***
## Month             0.46427    0.06229   7.454 1.04e-13 ***
## SITE_LATITUDE     6.52098    0.35275  18.486  < 2e-16 ***
## SITE_LONGITUDE   -0.49563    0.09850  -5.032 5.01e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 12.6 on 5771 degrees of freedom
##   (3200 observations deleted due to missingness)
## Multiple R-squared:  0.1926, Adjusted R-squared:  0.192
## F-statistic: 344.2 on 4 and 5771 DF,  p-value: < 2.2e-16
```