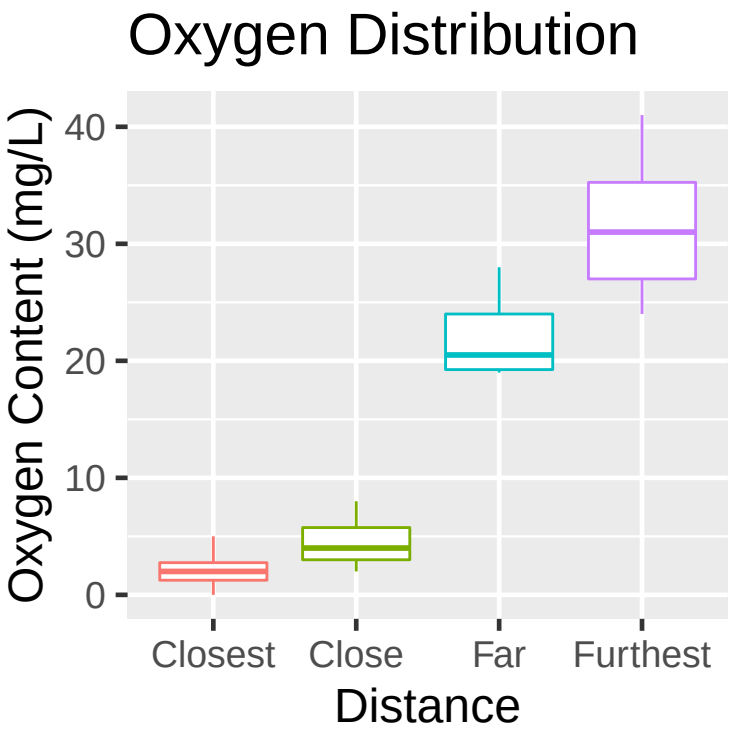
Homework Three

Josh Virene

Problem 1

Problem 1(a)



This graph showing distribution of oxygen by distance is a graphical representation of the information found in ANOVA. The spread of data around the mean within each factor (distance to the Mississippi) is the MS(error). The difference in means across factors is the MS(treatment).

Problem 1(b)

```
##
## Call:
## lm(formula = oxygen ~ factor, data = oxygendata)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -10.20  -1.60  -0.30   2.05   9.60
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
```

```
## (Intercept)    2.200      1.220   1.803   0.0798 .
## factor2        2.400      1.726   1.391   0.1728
## factor3       19.000      1.726  11.011 4.46e-13 ***
## factor4       29.200      1.726  16.921  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.859 on 36 degrees of freedom
## Multiple R-squared:  0.9153, Adjusted R-squared:  0.9083
## F-statistic: 129.7 on 3 and 36 DF,  p-value: < 2.2e-16
```

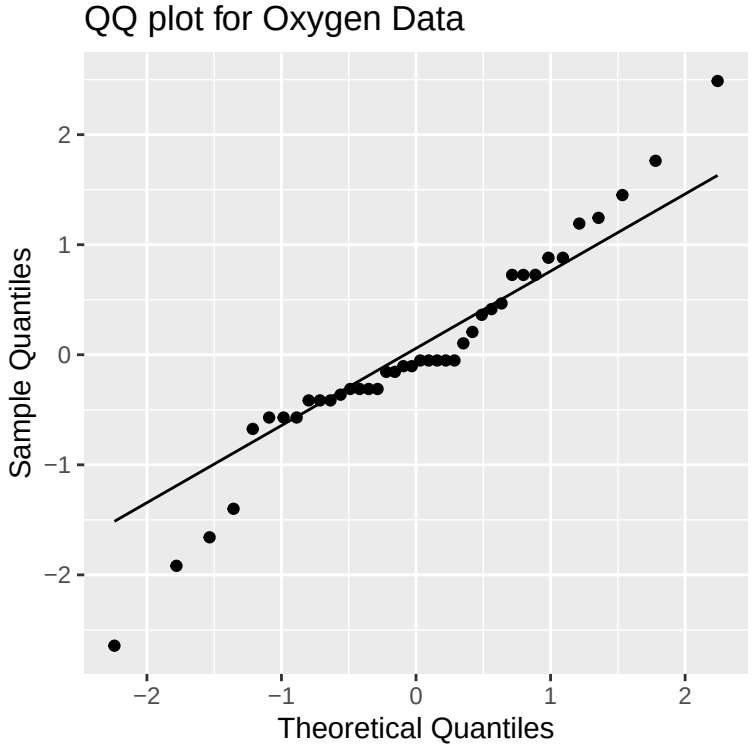
```
## Analysis of Variance Table
##
## Response: oxygen
##           Df Sum Sq Mean Sq F value    Pr(>F)
## factor     3 5793.1 1931.03   129.7 < 2.2e-16 ***
## Residuals 36  536.0   14.89
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

P value = 2.2e-16 Looking at the plot above, and the p value, it is clear we will reject the null hypothesis that all group population means are equal because the mean values across factors are very different, and the p value is extremely small.

Problem 1(c)

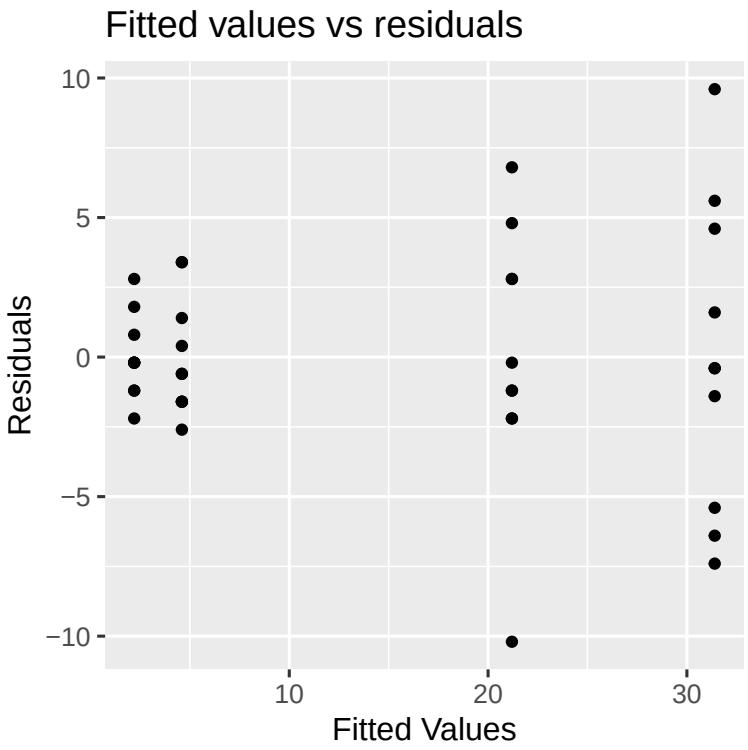
$H_o : \mu_1 = \mu_2 = ... = \mu_k$

Problem 1(d)



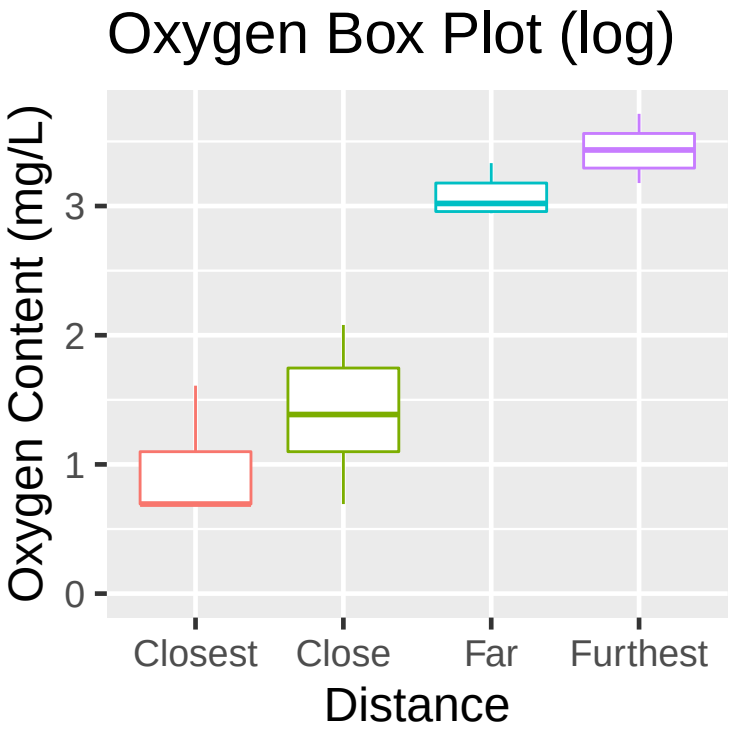
QQ plot: The QQ plot assesses the assumption that the error and residuals come from a normal distribution. If the points of the QQ plot are on the 1-1 line, this means that the assumption is met, otherwise it is violated. This plot looks like it violates the assumption, and in fact the tails appear to be heavy since the data are spread far from the 1-1 line in both directions.

Problem 1(d- continued)



Fitted values vs. Residual: The fitted values vs. residual plot assesses equality of variance. If the assumption is met, the variance of residuals will be roughly equal across groups, otherwise the assumption is violated. The data on this plot has a funnel shape; residual variance increases with fitted values so the assumption is violated.

Problem 1(e)



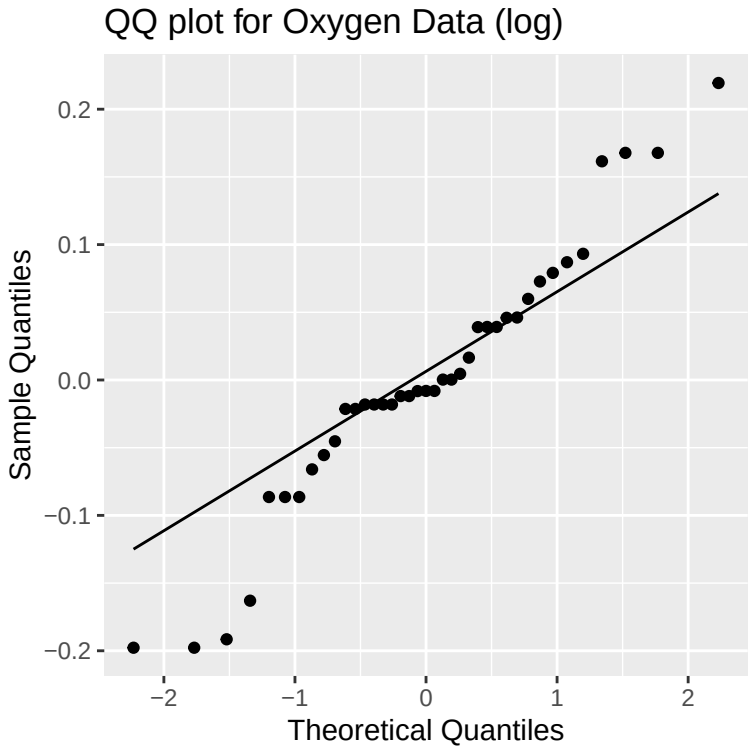
The distribution of means across factors is closer together in the log transformation compared to their distribution in the original question. The within group distribution is similar in the log transformation compared to the normal. Looking at the box plots in the log transformation, distance appears to a smaller impact on the mean value for dissolved oxygen content compared to the original form.

Problem 1(f)

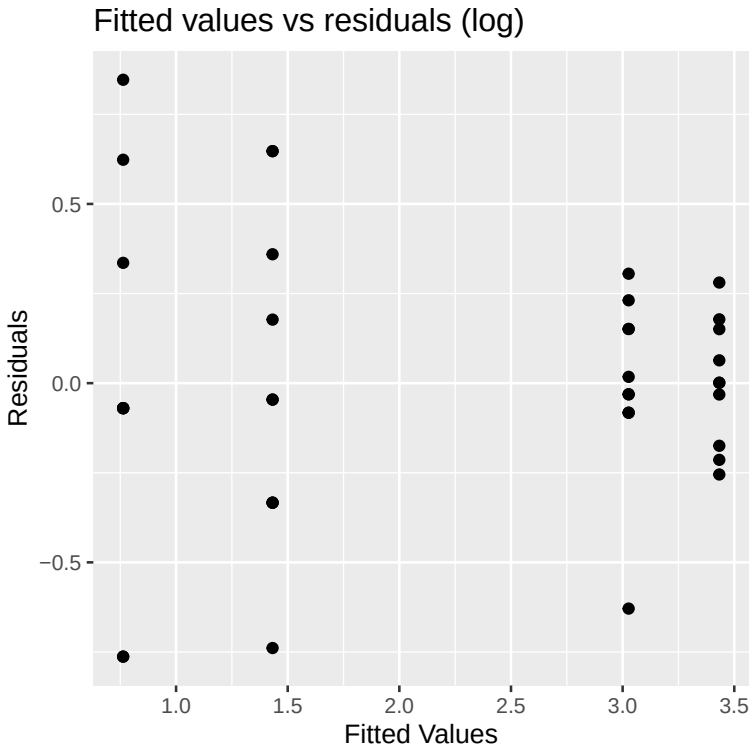
```
##
## Call:
## lm(formula = logoxygen ~ factor, data = oxygendata2)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.76299 -0.12863 -0.03118  0.17766  0.84645
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   0.7630     0.1281   5.957 8.80e-07 ***
## factor2       0.6692     0.1765   3.790  0.00057 ***
## factor3       2.2639     0.1765  12.823 8.73e-15 ***
## factor4       2.6699     0.1765  15.123  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3842 on 35 degrees of freedom
## Multiple R-squared:    0.9,  Adjusted R-squared:  0.8914
## F-statistic:   105 on 3 and 35 DF,  p-value: < 2.2e-16

## Analysis of Variance Table
```

```
##
## Response: logoxygen
##           Df Sum Sq Mean Sq F value    Pr(>F)
## factor     3 46.518 15.5061  105.02 < 2.2e-16 ***
## Residuals 35  5.168  0.1476
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```



QQ plot: The QQ plot assesses the assumption that the error and residuals come from a normal distribution. If the points of the QQ plot are on the 1-1 line, this means that the assumption is met, otherwise it is violated. This plot looks like it violates the assumption, and in fact the tails appear to be heavy since the data are spread far from the 1-1 line in both directions. This plot appears to be a more significant violation of the assumption compared to the original model because the points have more extreme deviations from the 1-1 line.



Fitted values vs. Residual: The theoretical vs. residual plot assesses equality of variance. If the assumption is met, the variance of residuals will be roughly equal across groups, otherwise the assumption is violated. The data on this plot has a funnel shape in the opposite direction compared to the original model- residual variance decreases with fitted values so the assumption is still violated. **Single factor ANOVA model:** Doing a log transformation on the data changes the interpretation of the model parameters. In the original model, the table shows additive differences, whereas in this log transformed model, they are multiplicative. The advantage is that multiplicative changes are more interpretable.

Problem 1(g)

The data in the original model appear to do a better job of satisfying the ANOVA assumptions because the plots show violations that are less extreme for the QQplot, and about the same for the residual vs. fitted plot. Despite this, neither plot is good in terms of meeting the ANOVA assumptions, both diagnostic plots exhibit serious violations.

Problem 1(h)

```
##
##  Kruskal-Wallis rank sum test
##
## data:  logoxygen by factor
## Kruskal-Wallis chi-squared = 32.927, df = 3, p-value = 3.337e-07
```

Test statistic: 32.927 P value: 3.337e-07

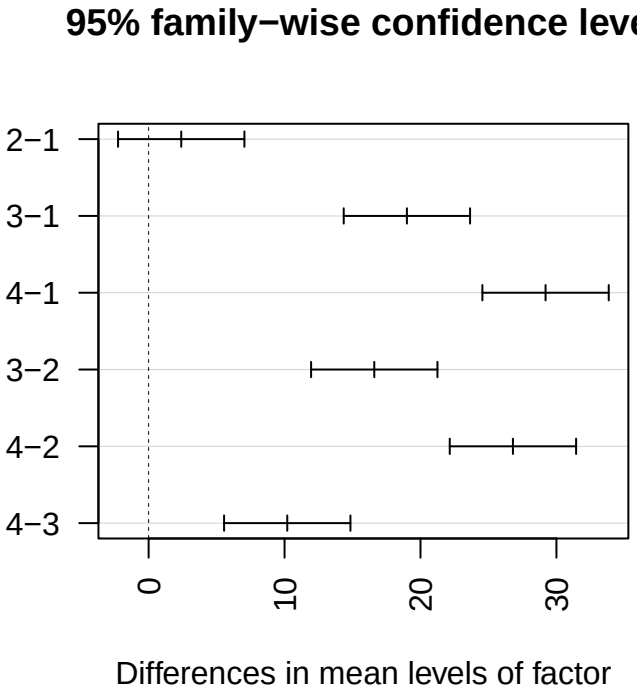
Problem 1(i)

When conducting a Kruskall-Wallace test, this is testing for equality in medians whereas the ANOVA test which tests equality in means. Another important consideration is that Kruskal-Wallis is nonparametric, meaning that it does not make assumptions about the distribution of errors, which can cause issues since it lowers power.

Problem 1(j)

```
##              p.raw  p.Bon p.Holm  p.fdr
## (Intercept) 0.0798 0.3191 0.1595 0.1064
## factor2     0.1728 0.6913 0.1728 0.1728
## factor3     0.0000 0.0000 0.0000 0.0000
## factor4     0.0000 0.0000 0.0000 0.0000
```

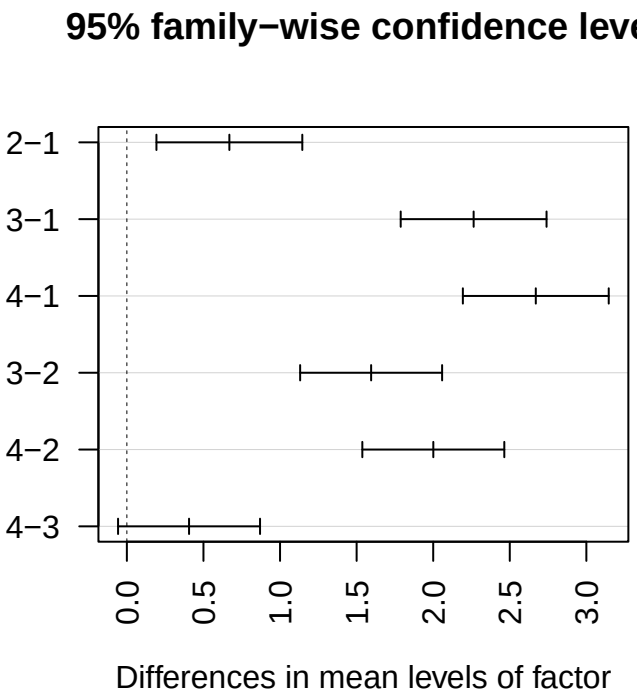
```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = oxygen ~ factor, data = oxygendata)
##
## $factor
##     diff       lwr       upr     p adj
## 2-1  2.4 -2.247497  7.047497 0.5131400
## 3-1 19.0 14.352503 23.647497 0.0000000
## 4-1 29.2 24.552503 33.847497 0.0000000
## 3-2 16.6 11.952503 21.247497 0.0000000
## 4-2 26.8 22.152503 31.447497 0.0000000
## 4-3 10.2  5.552503 14.847497 0.0000053
```




```
##             p.raw  p.Bon p.Holm p.fdr
## (Intercept) 0e+00 0.0000  0e+00 0e+00
## factor2     6e-04 0.0023  6e-04 6e-04
## factor3     0e+00 0.0000  0e+00 0e+00
```

```
## factor4     0e+00 0.0000  0e+00 0e+00
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = logoxygen ~ factor, data = oxygendata2)
##
## $factor
##          diff         lwr       upr     p adj
## 2-1 0.6691727  0.19303586 1.1453096 0.0030601
## 3-1 2.2639243  1.78778743 2.7400612 0.0000000
## 4-1 2.6698789  2.19374199 3.1460157 0.0000000
## 3-2 1.5947516  1.13131397 2.0581892 0.0000000
## 4-2 2.0007061  1.53726852 2.4641437 0.0000000
## 4-3 0.4059546 -0.05748305 0.8693922 0.1034719
```



In the original model, the confidence interval for the group two and group one pairwise comparison contains zero and therefore, this pairwise comparison is not significant at the 95% confidence level. All of the other pairwise comparisons are statistically significant.
In the log transformation model, the confidence interval for the group four and group three pairwise comparison contains zero and therefore, this pairwise comparison is not significant at the 95% confidence level. All of the other pairwise comparisons are statistically significant, though the groups two and one confidence interval is close to containing zero. The important difference between these two models is that the log transformation has changed which pairwise comparisons are statistically signficant and which are not.

Problem 1(k)

In choosing between these two models, I would pick the first model with raw oxygen because the ANOVA assumption violations are less extreme. Though, both models violate the ANOVA assumptions heavily, so neither is ideal. A reason for choosing the log transformed model instead of the original would be that multiplicative changes in the factor may be more interpretable than the additive interpretations of the original

model.

Problem 2

Problem 2(a)

i. This line of code generates a random normal variable with sample size N- defined in another line of code as 20, a mean of 20, and a standard deviation of 3.
ii. This tells R to treat variables A, B, and C as factors. Another important component in this is that these are being incorporated into the 'for loop' so R is computing the P values for each pairwise comparison adjustment across these factor through the assigned number of reps, which is 1000.
iii. This is telling R to create a data frame that reports each type of p values as assigned under their respective correction method.
iv. This line of code tells R to sum all of the p values less than 0.05. We multiply this by 1 at the end so that R can interpret values as true or false.

```
##   pvals.raw pvals.bon pvals.holm pvals.fdr
## 1     0.976     0.947      0.948      0.95
```

Any multiple comparison adjustment will reduce the power of the tests. The reasoning for this is that power is defined as the probability of rejecting $H_0|H_0$ is false. Implementing an adjustment method makes it harder to reject H_0, we get more type II errors and lower power. For Bonferroni, this method is the most conservative, as in, it drastically widens the confidence intervals, and overall decreases the p value threshold for rejecting H_0 so it is incredibly hard to reject H_0, which explains why this measure has the lowest power. The Holm-Bonferroni method is less conservative than the Bonferroni method, so power is slightly higher. Last, the false discovery rate takes a different approach. This method ensures that no more than 5% of rejected null hypotheses are type I errors. These are less conservative than methods that control family wise type I error rate and power will be higher.

```
##   pvals.raw pvals.bon pvals.holm pvals.fdr
## 1      0.54     0.386       0.48      0.51
```

Changing whichpair decreases power. What this change does is it changes the pairwise comparison for which we are estimating power, so the results differ based on the pairwise comparison. Looking at this code, it makes sense that the second pairwise comparison has higher power than the first because the difference in means is larger for this pairwise comparison than the first (delta = 2 versus delta = 1).

Problem 3

Multiple comparisons adjustments are important because they ensure that we retain at least 95% coverage rate in the data analysis. If we did not make these comparisons, this rate would be much lower. To illustrate this, assume we are running an experiment on 15 different groups. This means that there should be 15 different test statistics, all of which are evaluated and either rejected or accepted at the 95% confidence level. For individual tests, the coverage rate is 95% but in ANOVA, we are comparing all of these against eachother. In the case of an ANOVA with 15 different groups, the overall coverage rate is $1 - 0.95^{15} = 0.536$ for the entire ANOVA model. The multiple comparisons adjustments account for this, and ensure that a sufficient coverage rate is retained. The only drawback with these correction approaches is that there is a price to be paid in the form of widening confidence intervals; estimates can be very conservative in how they correct.

Appendix

```
library(knitr)
#install the tidyverse library (do this once)
#install.packages("tidyverse")
library(tidyverse)
knitr::opts_chunk$set(echo = FALSE, message = FALSE, warning = FALSE, fig.width = 4, fig.height = 4, ti
#code for prob1a
#load data
oxygendata <- read_csv("oxygen.csv")
view(oxygendata)
#convert distance to a factor
oxygendata$factor = as.factor(oxygendata$distance)

ggplot(oxygendata, aes(x=factor , y=oxygen)) +
geom_boxplot(aes(color = factor), outlier.colour = NA, position = "dodge") +
ylab("Oxygen Content (mg/L)") +
ggtitle("Oxygen Distribution") +
scale_x_discrete(name="Distance",breaks=c("1","2","3","4"),
labels=c("Closest","Close","Far","Furthest")) +
theme_gray(base_size = 18) +
theme(legend.position = "none")
#code for prob1b
oxygenlm1 <- lm(oxygen ~ factor, data = oxygendata)
summary(oxygenlm1)
anova(oxygenlm1)
#code for prob1c
#code for prob1d

#plot one- the QQplot
oxygenlm1 <- lm(oxygen ~ factor, data = oxygendata)
mse = anova(oxygenlm1)$"Mean Sq"[2]
scaledres = oxygenlm1$residuals/sqrt(mse)
s.df = data.frame(scaled = scaledres)
ggplot(s.df,aes(sample=scaled)) +
stat_qq() +
stat_qq_line() +
xlab("Theoretical Quantiles") +
ylab("Sample Quantiles") +
ggtitle("QQ plot for Oxygen Data")
#code for prob1dcontinued
#plot two- the residuals vs. fitted values plot
# Create data frame for ggplot
lm.df = data.frame(
  fitted = oxygenlm1$fitted.values,
  residuals = oxygenlm1$residuals,
  group = as.factor(oxygendata$distance))
#plot fitted vs residuals
ggplot(lm.df,aes(x=fitted,y=residuals))+
  geom_point() +
ggtitle("Fitted values vs residuals") +
xlab("Fitted Values") + ylab("Residuals") +
theme_gray(base_size = 12)
#code for prob1e
```

```
#creating the dataset:
oxygendata$logoxygen = log(oxygendata$oxygen)
view(oxygendata)
#removing the unwanted measurement
oxygendata2 <- oxygendata[-c(9), ]
view(oxygendata2)

#making the boxplot
ggplot(oxygendata2, aes(x=factor , y=logoxygen)) +
geom_boxplot(aes(color = factor), outlier.colour = NA, position = "dodge") +
ylab("Oxygen Content (mg/L)") +
ggtitle("Oxygen Box Plot (log)") +
scale_x_discrete(name="Distance",breaks=c("1","2","3","4"),
labels=c("Closest","Close","Far","Furthest")) +
theme_gray(base_size = 18) +
theme(legend.position = "none")
#code for prob1f

#setting up the anova
oxygenlm2 <- lm(logoxygen ~ factor, data = oxygendata2)
summary(oxygenlm2)
anova(oxygenlm2)
#code for prob1fcontd

#Creating the QQplot
oxygenlm2 <- lm(logoxygen ~ factor, data = oxygendata2)
mse2 = anova(oxygenlm2)$"Mean Sq"[2]
scaledres = oxygenlm2$residuals/sqrt(mse)
s.df2 = data.frame(scaled = scaledres)
ggplot(s.df2,aes(sample=scaled)) +
stat_qq() +
stat_qq_line() +
xlab("Theoretical Quantiles") +
ylab("Sample Quantiles") +
ggtitle("QQ plot for Oxygen Data (log)")
#code for prob1fcontinued

#Creating the residuals vs. fitted plot
lm.df2 = data.frame(
  fitted = oxygenlm2$fitted.values,
  residuals = oxygenlm2$residuals,
  group = as.factor(oxygendata2$distance))
#plot fitted vs residuals
ggplot(lm.df2,aes(x=fitted,y=residuals))+
  geom_point() +
ggtitle("Fitted values vs residuals (log)") +
xlab("Fitted Values") + ylab("Residuals") +
theme_gray(base_size = 10)
#code for prob1g

#code for prob1h
KWtest <- kruskal.test(logoxygen ~ factor, data = oxygendata2)
KWtest
```

```r
#code for prob1i


##model one:

p = data.frame(p.raw = summary(oxygenlm1)$coefficients[,4])
p$p.Bon = p.adjust(p$p.raw, method = "bonferroni")
p$p.Holm = p.adjust(p$p.raw, method = "holm")
p$p.fdr = p.adjust(p$p.raw, method = "fdr")
round(p,4)

oxygen1.Tukey <- TukeyHSD(aov(oxygen ~ factor, data = oxygendata))
oxygen1.Tukey
plot(oxygen1.Tukey,las = 2)

#model two
p = data.frame(p.raw = summary(oxygenlm2)$coefficients[,4])
p$p.Bon = p.adjust(p$p.raw, method = "bonferroni")
p$p.Holm = p.adjust(p$p.raw, method = "holm")
p$p.fdr = p.adjust(p$p.raw, method = "fdr")
round(p,4)

oxygen2.Tukey <- TukeyHSD(aov(logoxygen ~ factor, data = oxygendata2))
oxygen2.Tukey
plot(oxygen2.Tukey, las = 2)


#code for prob1k

#code for prob2a
#code for prob2b
set.seed(342)
reps <- 1000
N <- 20
pvals.raw <- rep(0,reps)
pvals.bon <- rep(0,reps)
pvals.holm <- rep(0,reps)
pvals.fdr <- rep(0,reps)
whichpair <- 2
for(i in 1:reps){
y1 = rnorm(n=N,mean=20,sd=3)
y2 = rnorm(n=N,mean=18,sd=3)
y3 = rnorm(n=N,mean=16,sd=3)
y = c(y1,y2,y3)
x = as.factor(c(rep("A",N),rep("B",N),rep("C",N)))
pvals.raw[i] = pairwise.t.test(y,x,p.adj="none")$p.value[whichpair]
pvals.bon[i] = pairwise.t.test(y,x,p.adj="bonferroni")$p.value[whichpair]
pvals.holm[i] = pairwise.t.test(y,x,p.adj="holm")$p.value[whichpair]
pvals.fdr[i] = pairwise.t.test(y,x,p.adj="fdr")$p.value[whichpair]
}
pvals <- data.frame(pvals.raw,pvals.bon,pvals.holm,pvals.fdr)
power.results <- as.data.frame((pvals<0.05)*1) %>%
summarize(across(.fns=sum)/1000)
```

```
power.results
#code for prob2c
set.seed(342)
reps <- 1000
N <- 20
pvals.raw <- rep(0,reps)
pvals.bon <- rep(0,reps)
pvals.holm <- rep(0,reps)
pvals.fdr <- rep(0,reps)
whichpair <- 1
for(i in 1:reps){
y1 = rnorm(n=N,mean=20,sd=3)
y2 = rnorm(n=N,mean=18,sd=3)
y3 = rnorm(n=N,mean=16,sd=3)
y = c(y1,y2,y3)
x = as.factor(c(rep("A",N),rep("B",N),rep("C",N)))
pvals.raw[i] = pairwise.t.test(y,x,p.adj="none")$p.value[whichpair]
pvals.bon[i] = pairwise.t.test(y,x,p.adj="bonferroni")$p.value[whichpair]
pvals.holm[i] = pairwise.t.test(y,x,p.adj="holm")$p.value[whichpair]
pvals.fdr[i] = pairwise.t.test(y,x,p.adj="fdr")$p.value[whichpair]
}
pvals <- data.frame(pvals.raw,pvals.bon,pvals.holm,pvals.fdr)
power.results <- as.data.frame((pvals<0.05)*1) %>%
summarize(across(.fns=sum)/1000)
power.results
#code for prob3
```