STAT 230 - Project Report - Group 2B

Names: Dasha Asienga, Ella Rose, Diego Carias

Project Title: Determiners of Fish Count in Williamette River for the Year 2020

Introduction

Our project is interested in examining what factors affect fish counts for certain species of Salmon and some white fish at Willamette Falls in the year 2020.

We set out to explore this question because of the recent decline of salmon in the Pacific Northwest. Salmon are important for ecosystems, they carry nutrients from oceans up rivers, and they have cultural significance for many native American groups, among others, yet there have been no useful solutions for their evident decline. As such, we were interested in understanding their decline by analyzing which predictors were most useful for explaining fish counts, particularly salmon.

We have several variables: location, time of year, river flow, river temperature, maximum air temperature, an indicator variable for whether there was a full moon or not, and the fish count, which is our response variable. We also have a species variable. We will use both a multiple linear regression model and a two-way additive ANOVA model to answer our questions.

Read in the data

```
group2bdata <- read_csv("https://awagaman.people.amherst.edu/stat230/projectsF21/group2BdataF21.csv")
```

```
## Rows: 366 Columns: 9
```

```
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (6): date, location, time of year, river temperature (C), full moon, spe...
## dbl (2): max air temperature (C), fish count
```

```
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
spec(group2bdata)
```

```
## cols(
##   date = col_character(),
##   location = col_character(),
##   `time of year` = col_character(),
##   `flow(ft3/s)` = col_number(),
##   `river temperature (C)` = col_character(),
##   `max air temperature (C)` = col_double(),
##   `full moon` = col_character(),
##   `fish count` = col_double(),
##   species = col_character()
## )
```

Summary command on data set

```
group2bdata <- group2bdata %>%
  rename("river_temperature" = "river temperature (C)",
         "time_of_year" = "time of year",
         "flow" = `flow(ft3/s)`,
         "max_air_temperature" = `max air temperature (C)`,
```

```
        "full_moon" = `full moon`,
        "fish_count" = `fish count`) %>%
  filter(river_temperature != "12.9[4]") %>%
  filter(fish_count != "NA") %>%
  mutate(river_temperature = as.numeric(river_temperature))
```

```
## Warning in mask$eval_all_mutate(quo): NAs introduced by coercion
```

```
glimpse(group2bdata)
```

```
## Rows: 253
## Columns: 9
## $ date                <chr> "1/1", "1/3", "1/5", "1/7", "1/9", "1/11", "1/13",~
## $ location            <chr> "Williamette", "Williamette", "Williamette", "Will~
## $ time_of_year        <chr> "Winter", "Winter", "Winter", "Winter", "Winter", ~
## $ flow                <dbl> 15700, 28200, 29700, 41600, 53400, 59400, 86300, 8~
## $ river_temperature   <dbl> 6.7, 7.2, 7.9, 7.7, 7.7, 7.2, 7.1, 6.5, 5.9, 6.3, ~
## $ max_air_temperature <dbl> 10, 11, 7, 10, 7, 8, 3, 1, 6, 9, 9, 11, 12, 9, 9, ~
## $ full_moon           <chr> "N", "N", "N", "N", "N", "N", "N", "N", "N", "N", ~
## $ fish_count          <dbl> 128, 184, 236, 262, 278, 286, 293, 299, 299, 323, ~
## $ species             <chr> "Steelhead", "Steelhead", "Steelhead", "Steelhead"~
```

Data Codebook

Our variables are:

Variable 1 - `Fish Count`. It is quantitative. It describes the count of fish of a specific species of fish on specific days of the year.

Variable 2 - `Time of year`. It is qualitative. It describes the season of the year. It's a factor with 4 levels: fall, winter, spring, and summer.

Variable 3 - `Max Air Temperature`. It is quantitative. It describes the maximum air temperature recorded on that specific day.

Variable 4 - `Full Moon`. It is qualitative. It describes whether there was a full moon on that day. It's a factor with two levels: Y and N. Y means that there was a full moon and N means that there was no full moon.

Variable 5 - `River Flow`. It is quantitative. It describes the cubic feet per seconds flow of water in the river.

Variable 6 - `River Temperature`. It is quantitative. It describes the average river temperature on that day.

Variable 7 - `Species`. It is qualitative. It describes the species of fish of the largest count (count recorded) that day.

Variable 8 - `Location`. It is qualitative. It describes the location of the fish count. It's a factor with two levels: Willamette and Leaburg.

Analysis, Models, and Results

Our analysis took on 3 steps: looking at univariate analyses for each of the variables in our data set, looking at bivariate analyses for pairs of variables in our data set, and finally, conducting multivariate analysis in our final models.

Our univariate analysis helped us to identify what variables to re-express and our bivariate analysis highlighted interesting relationships between our variables, for example, discovering that we needed to proceed with two different data sets for both Willamette and Leaburg. We ended up with 2 final models: a multiple linear regression model with 4 variables for Willamette and a two-way additive ANOVA model for Willamette. Both have the square root of fish count as a predictor.

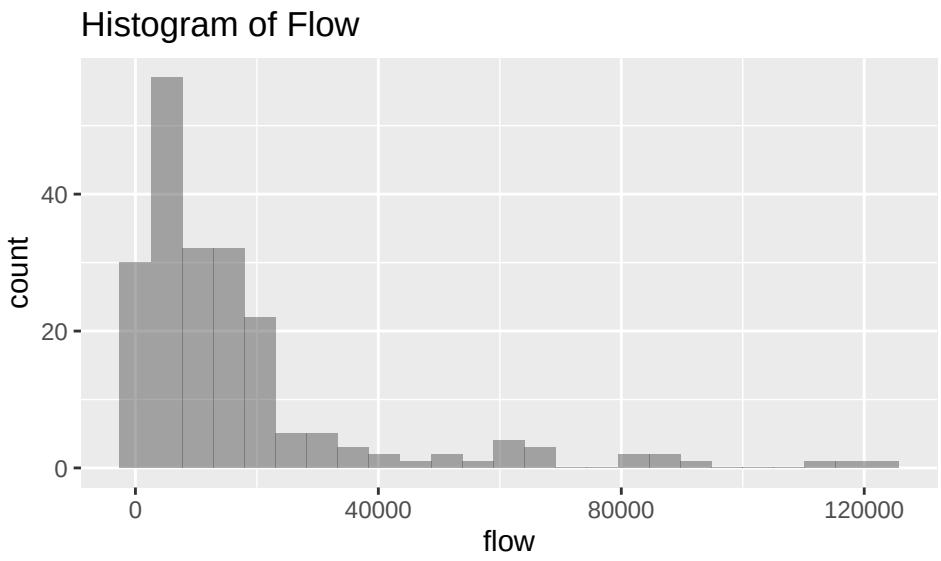
NB. Throughout the analysis, we comment out code chunks that don't need to be included in the report, but we leave them in as a comment just in case the reader wants to examine them closer.

Preliminary Univariate Analysis

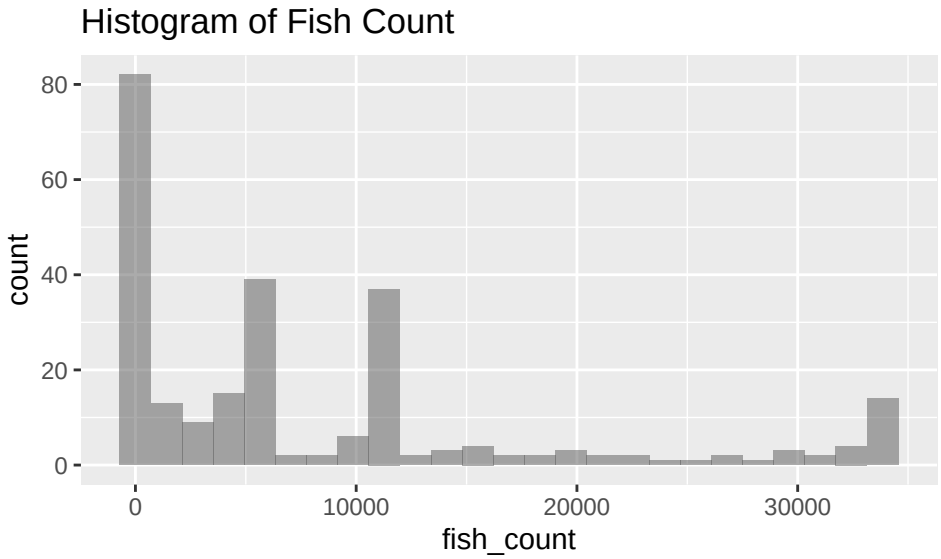
Quantitative Variables For our quantitative variables, we discovered that both the `flow` variable and the `fish count` variable had extreme right skews and needed transformation. The mean of the `flow` variable was 16698.21 while the median was 9420, and the mean of the `fish count` variable was 8248.383 while the median was 5299. The summary statistics for the other 2 quantitative variables are displayed in the summary table below. Those have a more normal distribution.

```
#for quantitative variables

gf_histogram(~ flow, data = group2bdata, title = "Histogram of Flow")
```

```
## Warning: Removed 46 rows containing non-finite values (stat_bin).
```

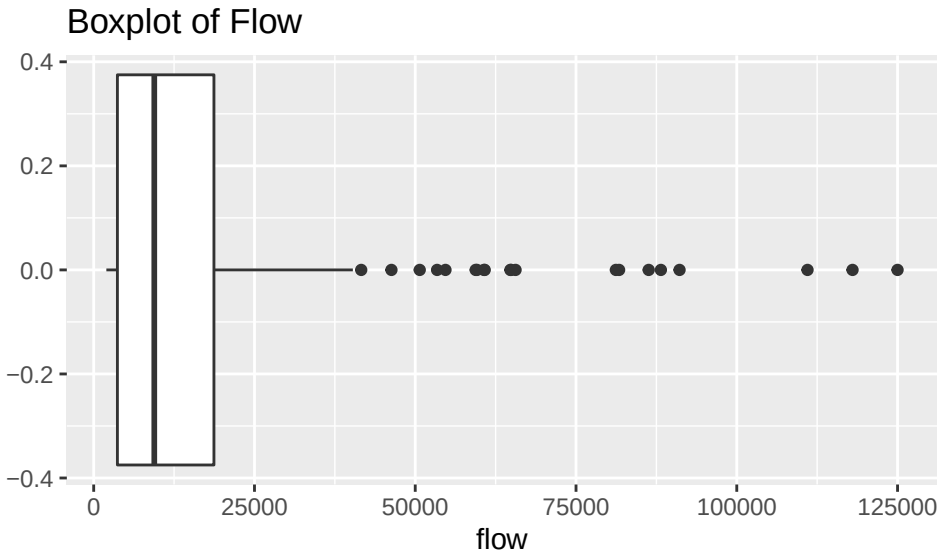



```
#gf_histogram(~ river_temperature, data = group2bdata)
#gf_histogram(~ max_air_temperature, data = group2bdata)
gf_histogram(~ fish_count, data = group2bdata, title = "Histogram of Fish Count")
```

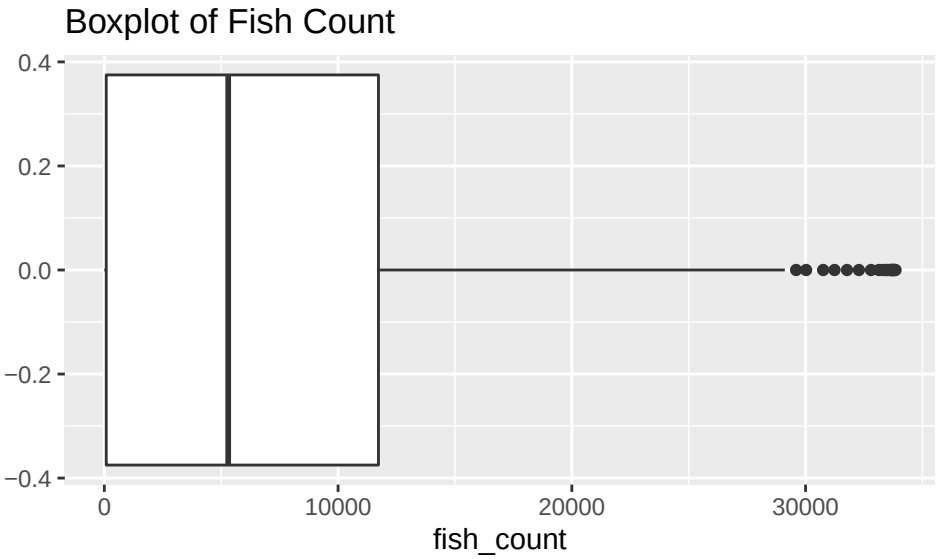



```
gf_boxplot(~ flow, data = group2bdata, title = "Boxplot of Flow")
```

```
## Warning: Removed 46 rows containing non-finite values (stat_boxplot).
```




```
#gf_boxplot(~ river_temperature, data = group2bdata)
#gf_boxplot(~ max_air_temperature, data = group2bdata)
gf_boxplot(~ fish_count, data = group2bdata, title = "Boxplot of Fish Count")
```




```
# gf_dens(~ flow, data = group2bdata)
# gf_dens(~ river_temperature, data = group2bdata)
# gf_dens(~ max_air_temperature, data = group2bdata)
# gf_dens(~ fish_count, data = group2bdata)

favstats(~ flow, data = group2bdata, title = "Summary of Flow")
```

```
##   min   Q1 median    Q3    max     mean       sd   n missing
##  1970 3685   9420 18700 125000 16698.21 21408.78 207      46
```

```
favstats(~ river_temperature, data = group2bdata, title = "Summary of River Temperature")
```

```
##  min Q1 median     Q3  max     mean      sd   n missing
##  5.9  9   13.5 16.325 24.3 13.36508 4.90183 252       1
```

```
favstats(~ max_air_temperature, data = group2bdata, title = "Summary of Maximum Air Temperature")
```

```
##  min Q1 median Q3 max     mean       sd   n missing
##    1 12     19 25  37 18.75099 8.132658 253       0
```

```
favstats(~ fish_count, data = group2bdata, title = "Summary of Fish Count")
```

```
##  min Q1 median    Q3   max     mean       sd   n missing
##    1 81   5299 11727 33848 8248.383 9876.244 253       0
```

Categorical Variables For our categorical variables, we see that the species and time of year are fairly balanced with a reasonably divided amount of data for each level. Location however, is not balanced as we have more that twice the amount of data for Willamette (183 observations) than we do for Leaburg (70 observations). The full moon exhibits even more extreme trends with 245 days with no full moon versus 8 days with a full moon. This makes sense as the moon rotates on a cycle and there can only be, at most, 12 full moons a year with most days not having a full moon.

```
#for categorical variables

tally(~ species, data = group2bdata)
```

```
## species
##        Ch Mark           Coho   Fall Chinook Spring Chinook      Steelhead
##             32             42             27             56             58
```

```
##     White Fish
##             38
```

```
tally(~ location, data = group2bdata)
```

```
## location
##     Leaburg Williamette
##          70         183
```

```
tally(~ time_of_year, data = group2bdata)
```

```
## time_of_year
##   Fall Spring Summer Winter
##     52     72     85     44
```

```
tally(~ full_moon, data = group2bdata)
```

```
## full_moon
##   N   Y
## 245   8
```

Summary of our Univariate Analysis `Flow`: It's extremely right-skewed. It has a lot of outliers on the higher end. There are 46 missing data points. The mean is much higher than the median, meaning that the outliers may have strong influence.

`River Temperature`: It's pretty normal, almost bimodal. The boxplot looks good. The mean and median are quite similar.

`Max Air Temperature`: This looks normal. The boxplot looks good. The mean and median are quite similar.

`Fish Count`: This is extremely right-skewed. This has quite a number of outliers on the higher end. The maximum is really high, which is pulling the mean towards it. This makes the mean different from the median.

`Species`: There are 6 species of fish recorded. The n for each species ranges from 27 to 58.

`Location`: There is more data for Willamette than for Leaburg. The data for Willamette is twice the amount that of Leaburg.

`Time of Year`: The count for time of year is pretty evenly distributed, but there are notably less winter days than there are summer days.

`Full Moon`: There were only 8 full moons recorded against 245, but that is expected.

Preliminary Multivariate Analysis In this section, we first examined scatter plots and side-by-side box plots to examine bivariate relationships that will later be useful for building our final models.

Bivariate Analysis Coloring the plots by location reveals very interesting characteristics. The numbers for Leaburg are always low and the numbers for Willamette vary by larger values, so the two distinct relationships are better seen. Because counts for Leaburg were always low and counts for Willamette were always high, we initially think that those two variables should be examined separately but we will revisit this.

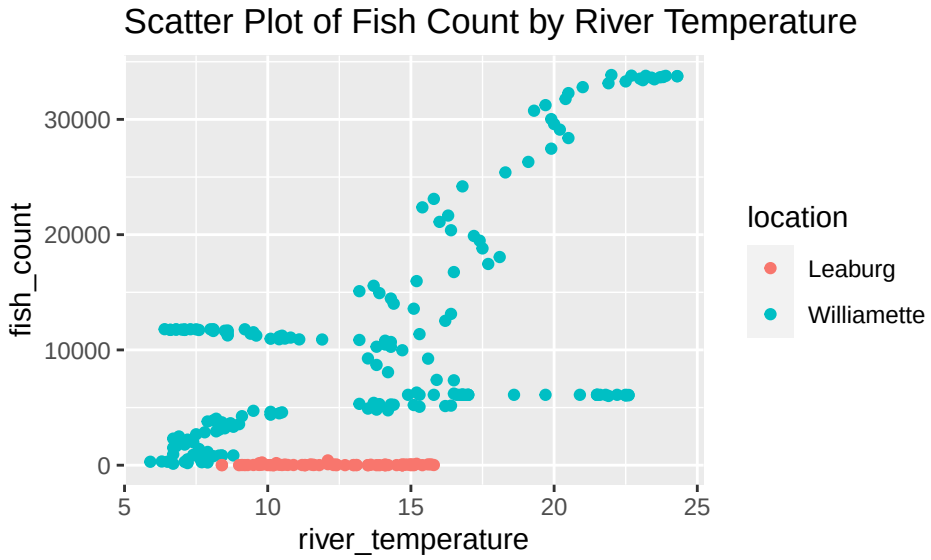
Maximum air temperature and river temperature have the strongest and most linear relationship of all the plots. They are highly correlated with a correlation coefficient of 0.758. All the predictors against `flow` have an interesting graph, further highlighting the need to re-express `flow`.

For the comparative box plots, we see an initial difference in means across all 4 categorical predictors. The difference for the 2 levels of the `full moon` variable is not as visible as it is for the other categorical predictors.

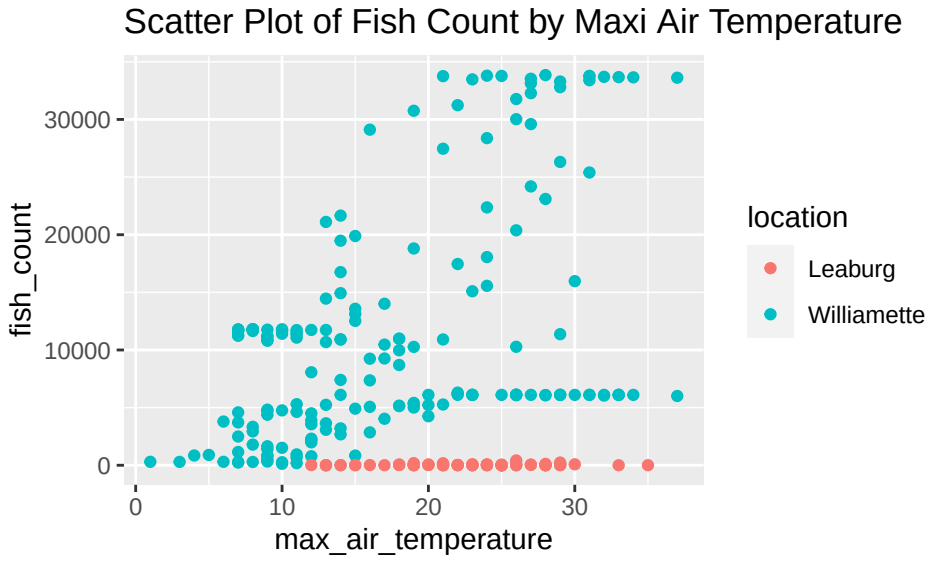
Lastly, the `ggpairs` command helps us visualize these bivariate relationships further.

```
gf_point(fish_count ~ river_temperature, color = ~ location, data = group2bdata,
         title = "Scatter Plot of Fish Count by River Temperature")
```

```
## Warning: Removed 1 rows containing missing values (geom_point).
```

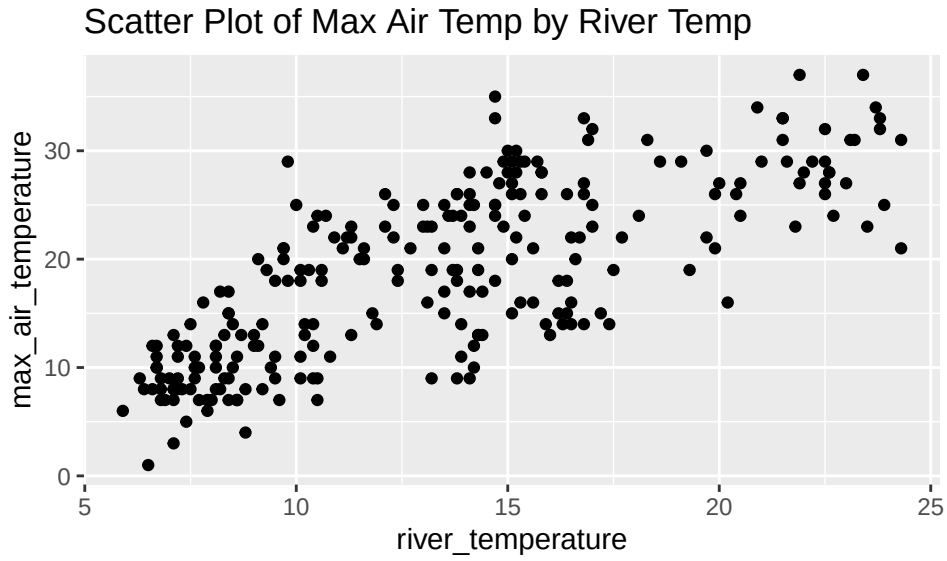



```
gf_point(fish_count ~ max_air_temperature, color = ~ location, data = group2bdata,
         title = "Scatter Plot of Fish Count by Maxi Air Temperature")
```



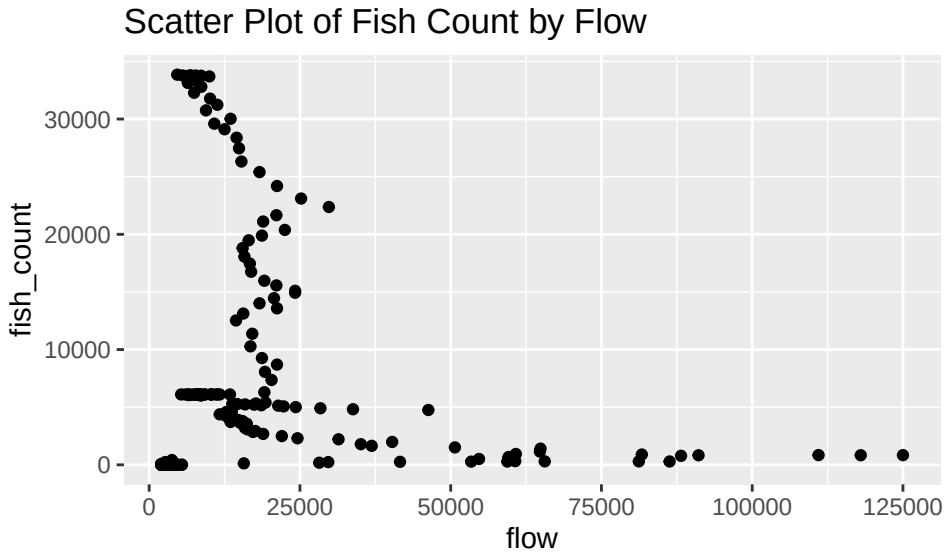

```
gf_point(max_air_temperature ~ river_temperature, data = group2bdata,
         title = "Scatter Plot of Max Air Temp by River Temp")
```

```
## Warning: Removed 1 rows containing missing values (geom_point).
```

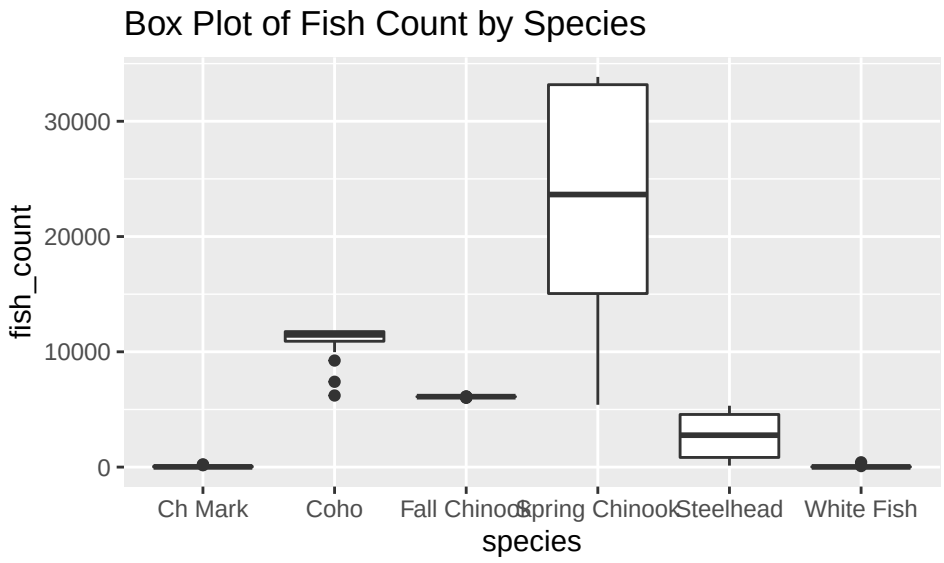



```
gf_point(fish_count ~ flow, data = group2bdata,
         title = "Scatter Plot of Fish Count by Flow")
```

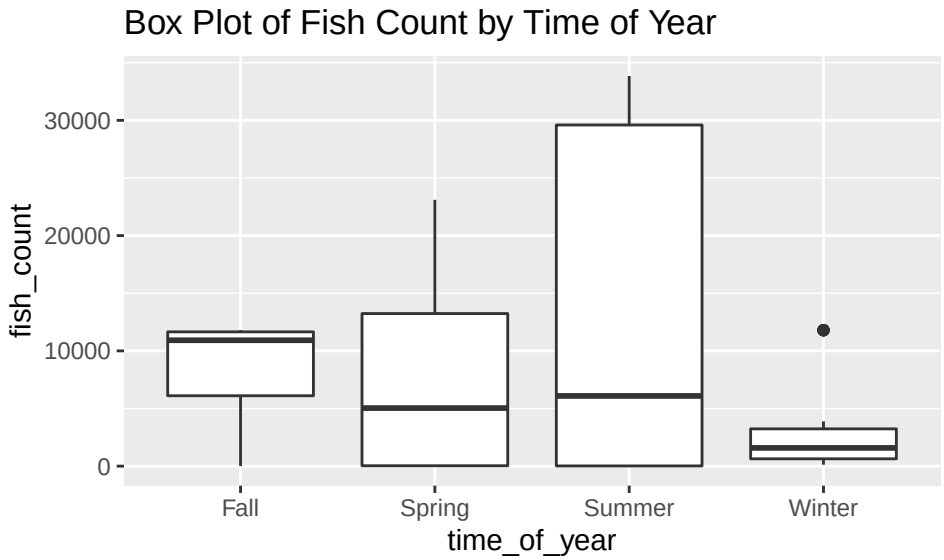
```
## Warning: Removed 46 rows containing missing values (geom_point).
```



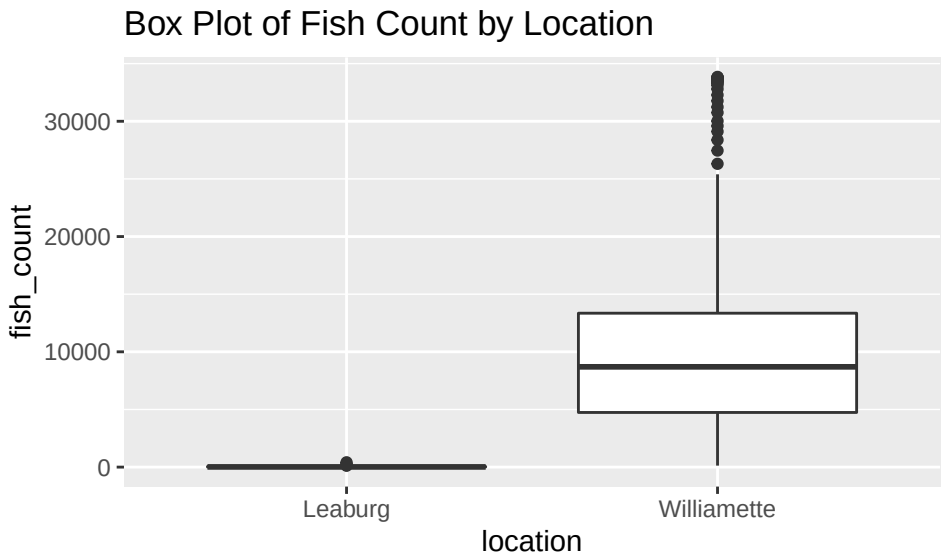

```
#gf_point(flow ~ river_temperature, color = ~ location, data = group2bdata)
#gf_point(max_air_temperature ~ flow, data = group2bdata)

gf_boxplot(fish_count ~ species, data = group2bdata,
           title = "Box Plot of Fish Count by Species")
```



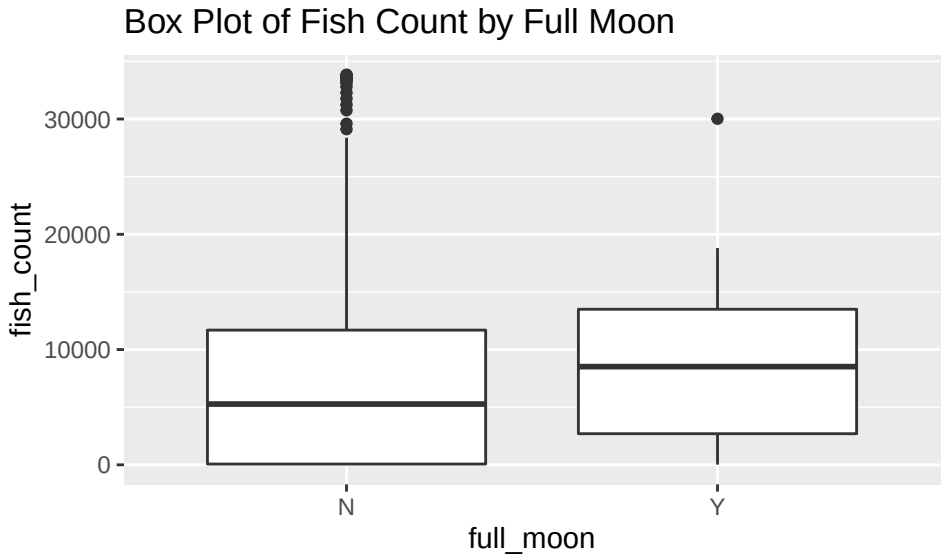

```
gf_boxplot(fish_count ~ time_of_year, data = group2bdata,
           title = "Box Plot of Fish Count by Time of Year")
```




```
gf_boxplot(fish_count ~ location, data = group2bdata,
           title = "Box Plot of Fish Count by Location")
```



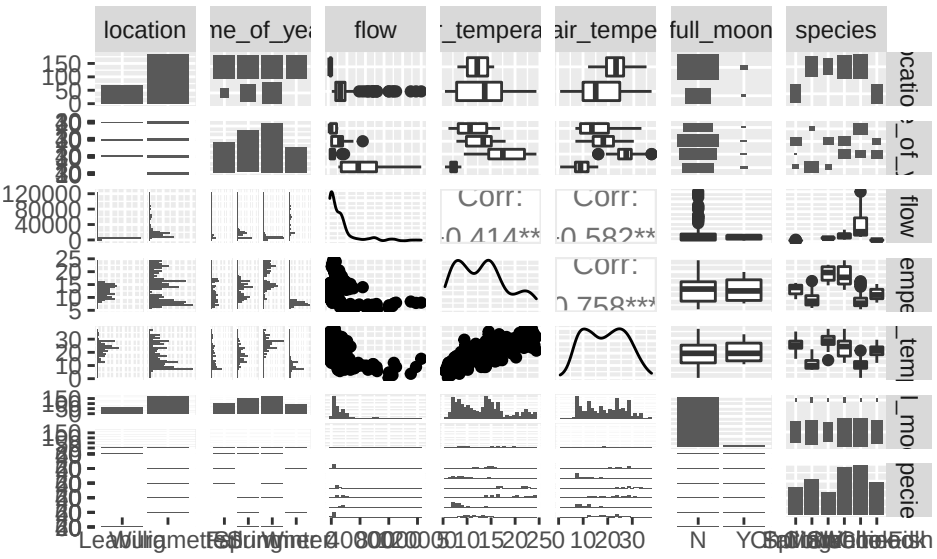

```
gf_boxplot(fish_count ~ full_moon, data = group2bdata,
           title = "Box Plot of Fish Count by Full Moon")
```




```
# Create data set without the date variable in order to compaire variables on ggpairs
group2bdata_nodate <- group2bdata %>%
  select(location, time_of_year, flow, river_temperature, max_air_temperature, full_moon, species)

ggpairs(group2bdata_nodate)
```

```
## Warning: Removed 46 rows containing non-finite values (stat_boxplot).
## Warning: Removed 1 rows containing non-finite values (stat_boxplot).
## Warning: Removed 46 rows containing non-finite values (stat_boxplot).
## Warning: Removed 1 rows containing non-finite values (stat_boxplot).
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## Warning: Removed 46 rows containing non-finite values (stat_bin).
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

```
## Warning: Removed 46 rows containing non-finite values (stat_bin).

## Warning: Removed 46 rows containing non-finite values (stat_density).

## Warning in ggally_statistic(data = data, mapping = mapping, na.rm = na.rm, :
## Removed 47 rows containing missing values

## Warning in ggally_statistic(data = data, mapping = mapping, na.rm = na.rm, :
## Removed 46 rows containing missing values

## Warning: Removed 46 rows containing non-finite values (stat_boxplot).

## Warning: Removed 46 rows containing non-finite values (stat_boxplot).

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

## Warning: Removed 1 rows containing non-finite values (stat_bin).

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

## Warning: Removed 1 rows containing non-finite values (stat_bin).

## Warning: Removed 47 rows containing missing values (geom_point).

## Warning: Removed 1 rows containing non-finite values (stat_density).

## Warning in ggally_statistic(data = data, mapping = mapping, na.rm = na.rm, :
## Removing 1 row that contained a missing value

## Warning: Removed 1 rows containing non-finite values (stat_boxplot).

## Warning: Removed 1 rows containing non-finite values (stat_boxplot).

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

## Warning: Removed 46 rows containing missing values (geom_point).

## Warning: Removed 1 rows containing missing values (geom_point).

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

## Warning: Removed 46 rows containing non-finite values (stat_bin).

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

## Warning: Removed 1 rows containing non-finite values (stat_bin).

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

## Warning: Removed 46 rows containing non-finite values (stat_bin).

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

## Warning: Removed 1 rows containing non-finite values (stat_bin).

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



Initial Models

Our initial kitchen sink model has all but 3 terms as significant on a 5% alpha level. The 3 non-significant predictors are maximum air temperature, which makes sense as this is highly correlated to river temperature with a correlation coefficient of 0.758. Additionally, full moon and whether it was summer were not considered significant. However, the p-value of summer is 0.067, which is quite close to 0.05. There is a remarkably high adjusted R-squared value of 0.9063. We proceed to fit more models, but our initial kitchen sink model is promising. Note that conditions have not been checked yet as well. Additionally, note that the kitchen sink model is still using a parallel slopes model for the `location` variable.

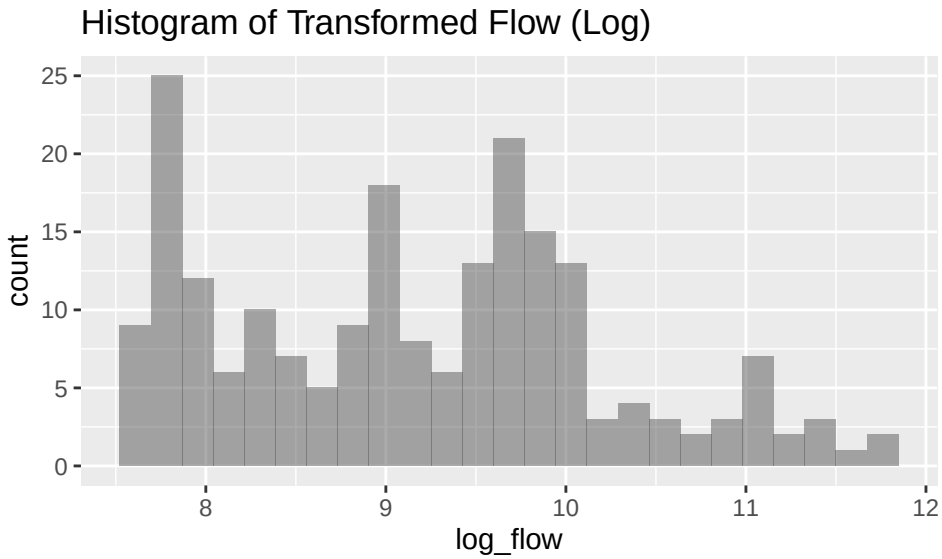
```
kitchensink <- lm(fish_count ~ location + time_of_year + flow + river_temperature +
             max_air_temperature + full_moon + species, data = group2bdata)
msummary(kitchensink)
```

```
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)           -9.545e+03  2.001e+03  -4.769 3.62e-06 ***
## locationWilliamette    1.157e+04  1.386e+03   8.344 1.32e-14 ***
## time_of_yearSpring    -7.245e+03  1.271e+03  -5.700 4.40e-08 ***
## time_of_yearSummer    -2.037e+03  1.107e+03  -1.840   0.0673 .
## time_of_yearWinter    -4.949e+03  1.919e+03  -2.579   0.0107 *
## flow                  -3.970e-02  1.630e-02  -2.436   0.0158 *
## river_temperature      8.207e+02  1.412e+02   5.813 2.49e-08 ***
## max_air_temperature    1.796e+01  5.667e+01   0.317   0.7516
## full_moonY             8.852e+02  1.368e+03   0.647   0.5183
## speciesFall Chinook   -1.058e+04  1.319e+03  -8.023 9.61e-14 ***
## speciesSpring Chinook  1.068e+04  1.044e+03  10.232  < 2e-16 ***
## speciesWhite Fish      5.113e+03  9.331e+02   5.480 1.31e-07 ***
##
## Residual standard error: 3269 on 194 degrees of freedom
##   (47 observations deleted due to missingness)
## Multiple R-squared:  0.9113, Adjusted R-squared:  0.9063
## F-statistic: 181.2 on 11 and 194 DF,  p-value: < 2.2e-16
```

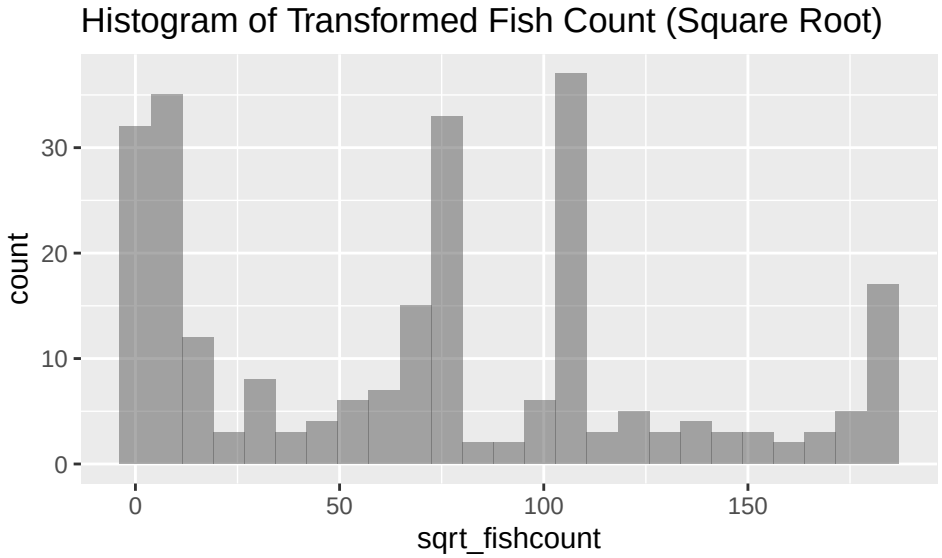
Re-expressing `flow` using log and `fish count` using square root was successful and the distributions for both variables are much more normal. We proceed to fit another kitchen sink model with these two transformed variables and obtain an even higher adjusted R-squared value of 0.9489 . This is very high but note, once again, that we still haven't checked conditions.

```r
# Transform the data
group2bdata_transformed <- group2bdata %>%
  mutate(log_flow = log(flow),
         sqrt_fishcount = sqrt(fish_count))

# New distributions of Transformed Data
gf_histogram(~log_flow, data = group2bdata_transformed,
             title = "Histogram of Transformed Flow (Log)")
```

```
## Warning: Removed 46 rows containing non-finite values (stat_bin).
```




```r
gf_histogram(~sqrt_fishcount, data = group2bdata_transformed,
             title = "Histogram of Transformed Fish Count (Square Root)")
```




```r
# Kitchensink model with transformed variables
kitchensink2 <- lm(sqrt_fishcount ~ location + time_of_year + log_flow + river_temperature +
             max_air_temperature + full_moon + species, data = group2bdata_transformed)
msummary(kitchensink2)
```

```
##                         Estimate Std. Error t value Pr(>|t|)
## (Intercept)              65.7149    24.2266   2.713 0.007278 **
## locationWilliamette     103.4125     6.9293  14.924  < 2e-16 ***
## time_of_yearSpring      -19.5097     5.3312  -3.660 0.000326 ***
## time_of_yearSummer       -8.4485     4.5875  -1.842 0.067056 .
## time_of_yearWinter      -29.3267     8.0000  -3.666 0.000318 ***
## log_flow                -11.8305     2.6409  -4.480 1.28e-05 ***
## river_temperature         2.5704     0.5948   4.322 2.47e-05 ***
## max_air_temperature       0.1955     0.2342   0.835 0.404810
## full_moonY                7.0060     5.6573   1.238 0.217069
## speciesFall Chinook     -33.5721     5.4607  -6.148 4.38e-09 ***
## speciesSpring Chinook    53.3189     4.3534  12.248  < 2e-16 ***
## speciesWhite Fish        15.8108     3.8634   4.092 6.26e-05 ***
##
## Residual standard error: 13.53 on 194 degrees of freedom
##   (47 observations deleted due to missingness)
## Multiple R-squared:  0.9517, Adjusted R-squared:  0.9489
## F-statistic: 347.4 on 11 and 194 DF,  p-value: < 2.2e-16
```

Before proceeding to our final models, we need to fit 2 separate models for Leaburg Dam and Willamette. The parallel slopes model we are currently considering does not match up to what the data and the plots are telling us.

The model for Leaburg has very low explanatory power with an adjusted R-squared of 0.1186. All the terms are not significant on a 5% alpha level as well, suggesting that the data we have for Leaburg doesn't exhibit any trends in explaining the square root of fish count. Even upon trying other pairings of predictors and using automated techniques to pick the best predictors, we still were not able to get a higher R-squared value.

The model for Willamette, however, is performing just as well as our full kitchen sink model with an adjusted R-squared of 0.947 and with many significant terms (the same terms as our kitchen sink model).

At this point, we decide to proceed with just the Willamette data.

```
# Data sets for the different locations:
leaburgdata <- group2bdata_transformed %>%
  filter(location == "Leaburg")

# Leaburg Kitchen Sink Model
leaburgmod <- lm(sqrt_fishcount ~ time_of_year + log_flow + river_temperature +
             max_air_temperature + full_moon + species, data = leaburgdata)
msummary(leaburgmod)
```

```
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)           17.7675    38.0777   0.467   0.6424
## time_of_yearSpring     1.1430     2.6421   0.433   0.6668
## time_of_yearSummer     0.7801     1.7237   0.453   0.6525
## log_flow              -1.2244     4.4779  -0.273   0.7854
## river_temperature     -0.7270     0.4425  -1.643   0.1055
## max_air_temperature    0.2526     0.1377   1.834   0.0715 .
## full_moonY             2.2075     2.6305   0.839   0.4046
## speciesWhite Fish     -1.2688     1.2645  -1.003   0.3196
##
## Residual standard error: 3.636 on 61 degrees of freedom
##   (1 observation deleted due to missingness)
## Multiple R-squared:  0.1186, Adjusted R-squared:  0.01743
## F-statistic: 1.172 on 7 and 61 DF,  p-value: 0.3319
```

```r
# Willamette Kitchen Sink Model
williamettedata <- group2bdata_transformed %>%
  filter(location == "Williamette")

willamettemod <- lm(sqrt_fishcount ~ time_of_year + log_flow + river_temperature +
             max_air_temperature + full_moon + species, data = williamettedata)
msummary(willamettemod)
```

```
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)           145.1087    27.2628   5.323 4.48e-07 ***
## time_of_yearSpring    -45.7703     6.8403  -6.691 6.37e-10 ***
## time_of_yearSummer    -11.0112     6.1733  -1.784 0.076864 .
## time_of_yearWinter    -58.2285     9.0035  -6.467 1.96e-09 ***
## log_flow              -11.6131     2.4081  -4.823 3.98e-06 ***
## river_temperature       2.1424     0.6105   3.509 0.000623 ***
## max_air_temperature     0.1529     0.2404   0.636 0.525771
## full_moonY              6.2414     6.3059   0.990 0.324163
## speciesSpring Chinook  98.7903     3.7852  26.099  < 2e-16 ***
## speciesSteelhead       54.3035     5.3885  10.078  < 2e-16 ***
##
## Residual standard error: 12.12 on 127 degrees of freedom
##   (46 observations deleted due to missingness)
## Multiple R-squared:  0.9505, Adjusted R-squared:  0.947
## F-statistic: 270.8 on 9 and 127 DF,  p-value: < 2.2e-16
```

Final Models

Multiple Linear Regression Model We use stepwise regression, an automated technique, to determine which predictors to include in our final multiple linear regression model.

```r
stepwise2 <- regsubsets(sqrt_fishcount ~ time_of_year + log_flow + river_temperature +
             max_air_temperature + full_moon + species, data = williamettedata, method = "seqrep", nbes

with(summary(stepwise2), data.frame(cp, outmat))
```

```
##                  cp time_of_yearSpring time_of_yearSummer time_of_yearWinter
## 1  ( 1 ) 556.525702
## 2  ( 1 ) 157.064433
## 3  ( 1 ) 109.976602
## 4  ( 1 )  48.520802                  *                                    *
## 5  ( 1 )  18.239709                  *                                    *
## 6  ( 1 ) 786.198838                  *                  *                 *
## 7  ( 1 ) 778.092810                  *                  *                 *
## 8  ( 1 )   8.404789                  *                  *                 *
##          log_flow river_temperature max_air_temperature full_moonY
## 1  ( 1 )
## 2  ( 1 )                         *
## 3  ( 1 )                         *
## 4  ( 1 )
## 5  ( 1 )        *
## 6  ( 1 )        *                *                  *
## 7  ( 1 )        *                *                  *         *
## 8  ( 1 )        *                *                             *
##          speciesSpring.Chinook speciesSteelhead
```

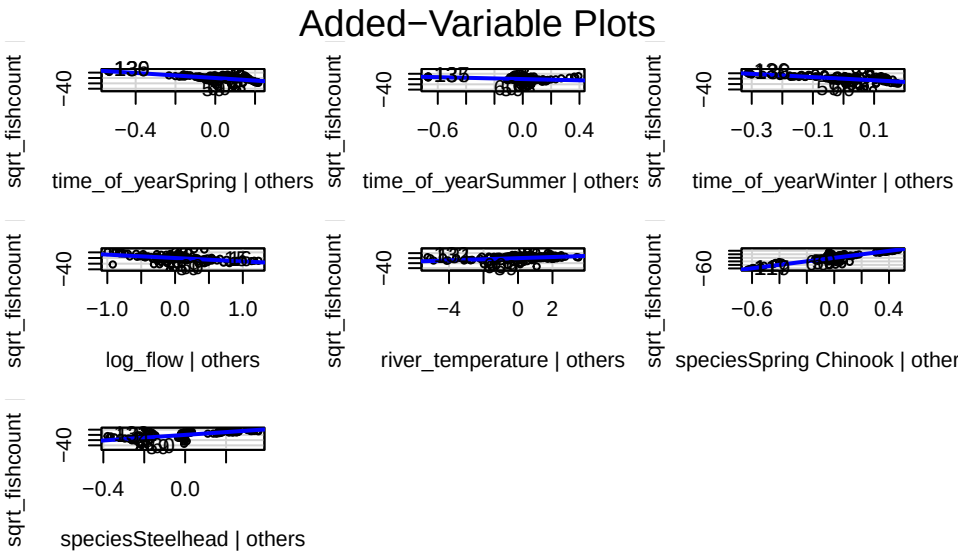
```
## 1  ( 1 )                    *
## 2  ( 1 )                    *
## 3  ( 1 )                    *                *
## 4  ( 1 )                    *                *
## 5  ( 1 )                    *                *
## 6  ( 1 )
## 7  ( 1 )
## 8  ( 1 )                    *                *
```

Using the stepwise regression technique and nested F tests, we landed at our final multiple linear regression model with 4 significant predictors: `time of year`, `river temperature`, `log of flow`, and `species`. The adjusted R-squared value is 0.9472. The model has an associated F value for 349.4 with a p-value of 0. Furthermore, the added variable plots showed that each of our predictors was useful in the final model after accounting for all other predictors.

```
finalmod <- lm(sqrt_fishcount ~ time_of_year + log_flow + river_temperature + species,
               data = williamettedata)
msummary(finalmod)
```

```
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)           151.3364    26.3701   5.739 6.46e-08 ***
## time_of_yearSpring    -47.1072     6.7386  -6.991 1.32e-10 ***
## time_of_yearSummer    -12.0132     6.0405  -1.989 0.048843 *
## time_of_yearWinter    -59.4591     8.9121  -6.672 6.74e-10 ***
## log_flow              -11.9316     2.3847  -5.003 1.80e-06 ***
## river_temperature       2.2418     0.5964   3.759 0.000257 ***
## speciesSpring Chinook  98.6685     3.7260  26.481  < 2e-16 ***
## speciesSteelhead       53.5035     5.2079  10.274  < 2e-16 ***
##
## Residual standard error: 12.09 on 129 degrees of freedom
##   (46 observations deleted due to missingness)
## Multiple R-squared:  0.9499, Adjusted R-squared:  0.9472
## F-statistic: 349.4 on 7 and 129 DF,  p-value: < 2.2e-16
```

```
car::avPlots(finalmod)
```

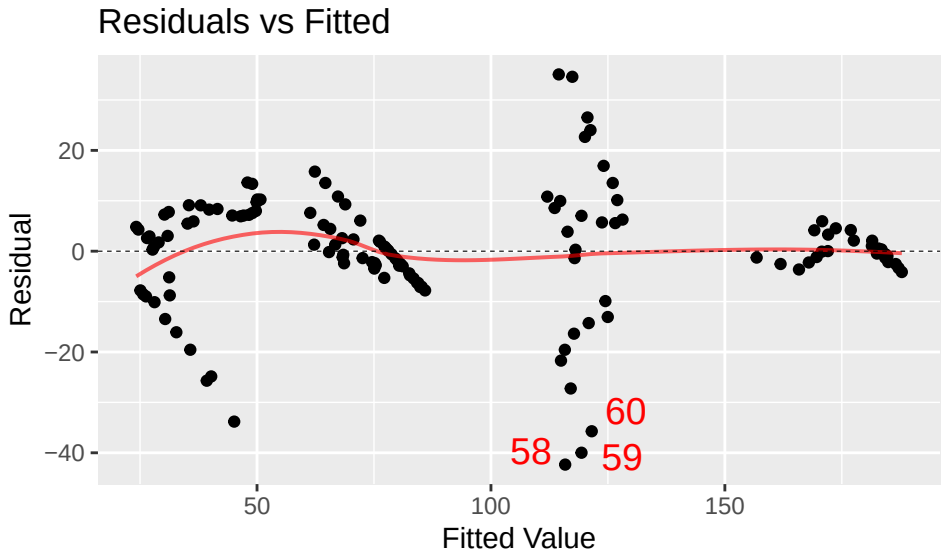


Checking MLR Diagnostics For the final MLR model, the conditions of normality and equal variance of errors were not met given their respective plots. In the Normal Q-Q plot, both tail ends deviate from the line.

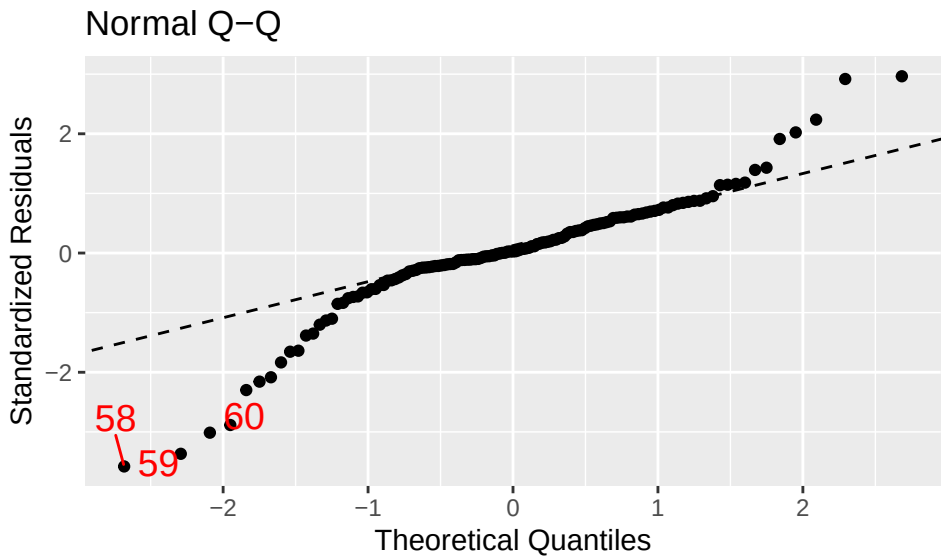
For the residuals vs. fitted plot, there is no equal spread. Afterwards, a plot to determine Cook' Distance for unusual points revealed that there are no substantial outliers for this model as no data point exceeded the .5 cutoff. Additionally, a cut-off value of .0656 for leverage was obtained in order to further explore potential outliers in a Residuals vs. Leverage plot and Cook's Distance Vs Leverage plot. As a result of this very high leverage cutoff, there are many points in the model that exceed the value, indicating potential outliers. In order to further investigate the usefulness of explanatory variables in predicting fish counts and check for multicollinearity, a variance inflation factor (VIF) value was obtained for all variables chosen in the model. The variables `time_of_year`, `log_flow`, `river_temperature`, and `species` do not have issues with multicollinearity as their VIF values were under 5.

```
#Final model diagnostics
mplot(finalmod, which = 1)
```

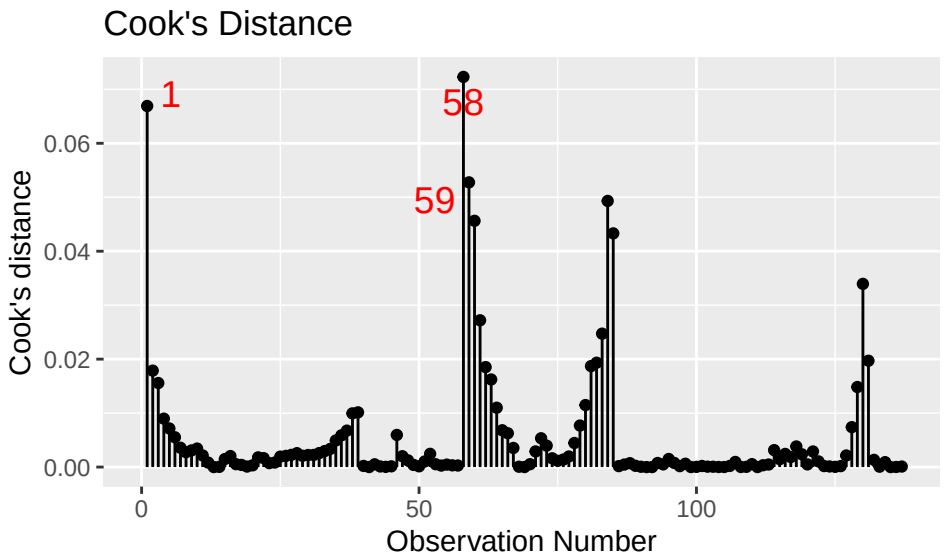
```
## `geom_smooth()` using formula 'y ~ x'
```




```
mplot(finalmod, which = 2)
```

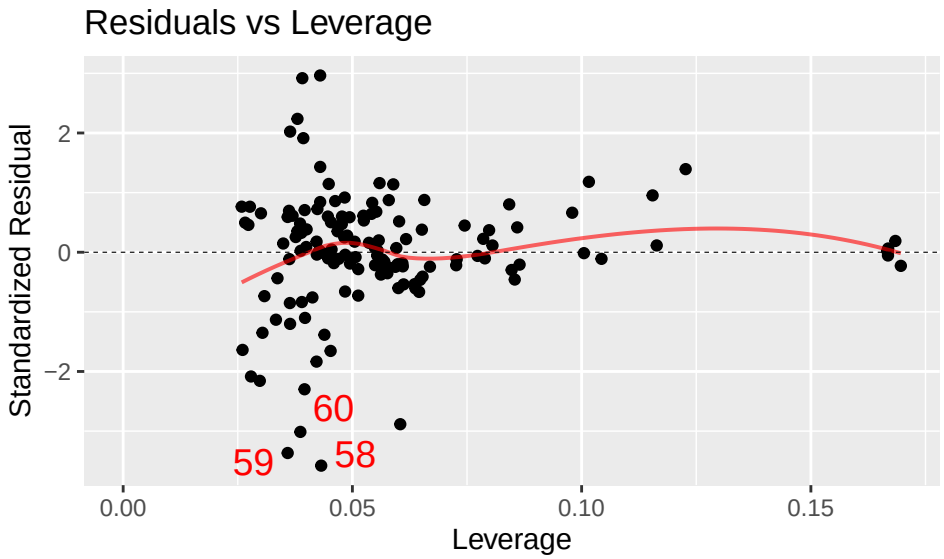



```
mplot(finalmod, which = 4)
```



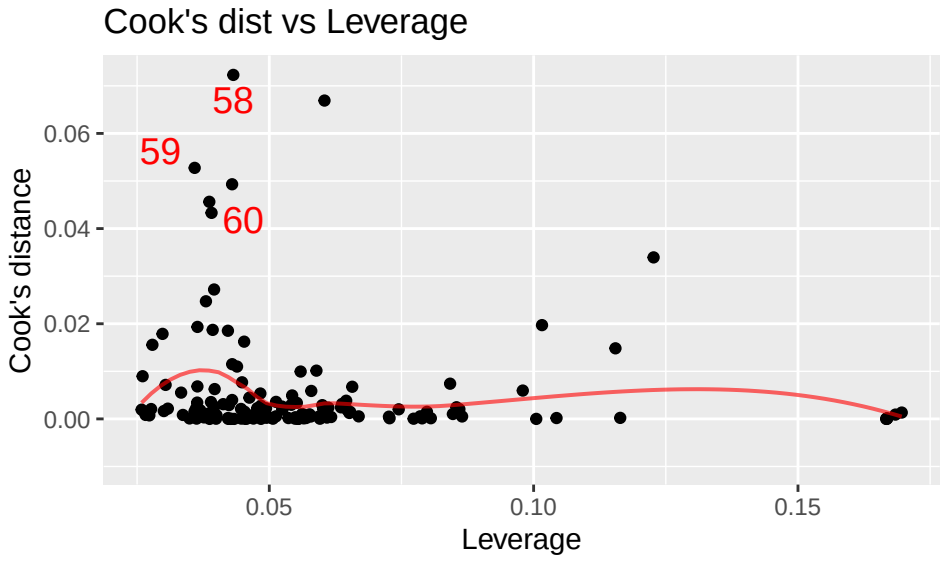

```
mplot(finalmod, which = 5)
```

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




```
mplot(finalmod, which = 6)
```

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




```
#Compute VIF
car::vif(finalmod)
```

```
##                         GVIF Df GVIF^(1/(2*Df))
## time_of_year      15.893080  3        1.585628
## log_flow           2.905659  1        1.704599
## river_temperature 10.570220  1        3.251187
## species            5.804379  2        1.552169
```

As seen, the conditions do not check out for our final multiple linear regression model. We will perform a randomization test for F later.

ANOVA The stepwise regression automated technique found that `species` was the single best predictor for the square root of fish count. This motivated the decision to have a one-way ANOVA as the next model in our analysis. An initial one-way ANOVA with just species as a predictor shows that `species` is a significant predictor for the square root of fish count. As a predictor, it accounts for 77.6% of the variance in the square root of fish count. The ANOVA output has an F-value of 206.8 with a p-value of 0.

```
anova1 <- lm(sqrt_fishcount ~ species, data = williamettedata)
msummary(anova1)
```

```
##                        Estimate Std. Error t value Pr(>|t|)
## (Intercept)             105.142      3.415  30.792  < 2e-16 ***
## speciesFall Chinook     -27.048      5.459  -4.955 1.67e-06 ***
## speciesSpring Chinook    43.476      4.517   9.625  < 2e-16 ***
## speciesSteelhead        -57.675      4.484 -12.863  < 2e-16 ***
##
## Residual standard error: 22.13 on 179 degrees of freedom
## Multiple R-squared:  0.7761, Adjusted R-squared:  0.7723
## F-statistic: 206.8 on 3 and 179 DF,  p-value: < 2.2e-16
```

```
anova(anova1)
```

```
## Analysis of Variance Table
##
## Response: sqrt_fishcount
##            Df Sum Sq Mean Sq F value    Pr(>F)
## species     3 303821  101274   206.8 < 2.2e-16 ***
```

```
## Residuals 179  87658     490
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

A nested F-Test for ANOVA revealed that `time of year` was useful in the model after accounting for the effects of `species`. None of the other categorical predictors, namely `full moon`, added more useful information.

Therefore, the final two-way additive ANOVA model had two predictors: `species`, which has an F-value of 412.8280 and an associated p-value of 0, and `time of year`, which has an F-value of 60.2376 and as associated p-value of 0.

This means that there is at least one difference in means of the square root of fish count detected for both categorical variables. In looking at Tukey's HSD, there are significant differences across all species in our data set. As for the time of year, there were no differences detected between spring and fall and between winter and spring. There were significant differences detected between all other seasons of the year.

```
anova2 <- lm(sqrt_fishcount ~ species + time_of_year, data = williamettedata)
anova(anova2)
```

```
## Analysis of Variance Table
##
## Response: sqrt_fishcount
##               Df Sum Sq Mean Sq F value    Pr(>F)
## species        3 303821  101274 411.399 < 2.2e-16 ***
## time_of_year   3  44332   14777  60.029 < 2.2e-16 ***
## Residuals    176  43326     246
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
TukeyHSD(anova2)
```

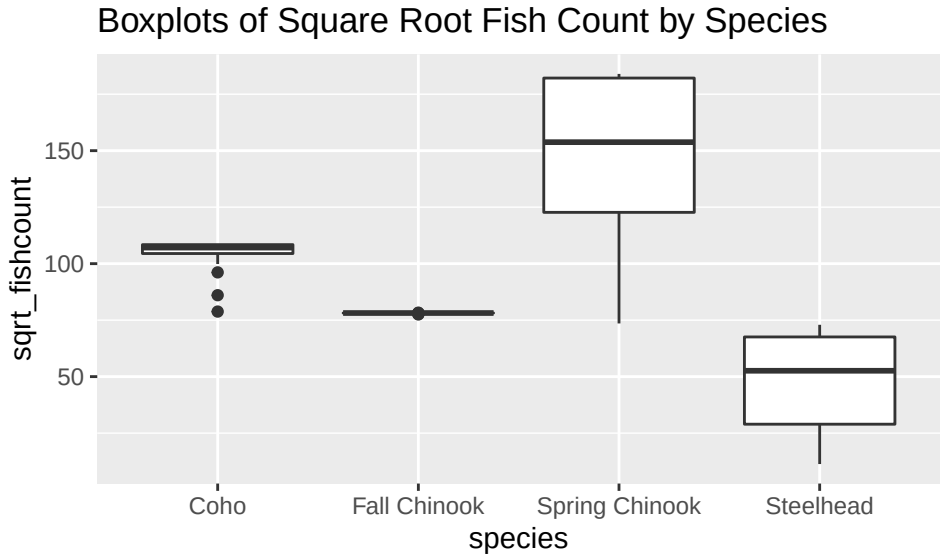
```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = x)
##
## $species
##                                   diff        lwr       upr p adj
## Fall Chinook-Coho            -27.04763  -37.08597 -17.00928     0
## Spring Chinook-Coho           43.47604   35.16915  51.78293     0
## Steelhead-Coho               -57.67487  -65.92015 -49.42959     0
## Spring Chinook-Fall Chinook   70.52366   60.98896  80.05837     0
## Steelhead-Fall Chinook       -30.62724  -40.10832 -21.14617     0
## Steelhead-Spring Chinook    -101.15091 -108.77501 -93.52681     0
##
## $time_of_year
##                      diff        lwr         upr     p adj
## Spring-Fall    -7.8289422 -16.316502   0.6586181 0.0822457
## Summer-Fall    17.5296641   9.042104  26.0172244 0.0000016
## Winter-Fall    -8.8172356 -17.445164  -0.1893068 0.0431133
## Summer-Spring  25.3586063  16.963809  33.7534036 0.0000000
## Winter-Spring  -0.9882934  -9.524985   7.5483978 0.9905628
## Winter-Summer -26.3468997 -34.883591 -17.8102085 0.0000000
```

Checking ANOVA Diagnostics The constant variance condition does not check out, as seen in the residual versus fitted plot, the comparative box plots, and the rule of 2 is not satisfied. There is no equal

variance. However, the normality condition does not look too concerning, as much as the points are not perfectly fitted along the line.

Because conditions do not check out, we will perform a randomization test for F.

```
gf_boxplot(sqrt_fishcount ~ species, data = williamettedata,
           title = "Boxplots of Square Root Fish Count by Species")
```

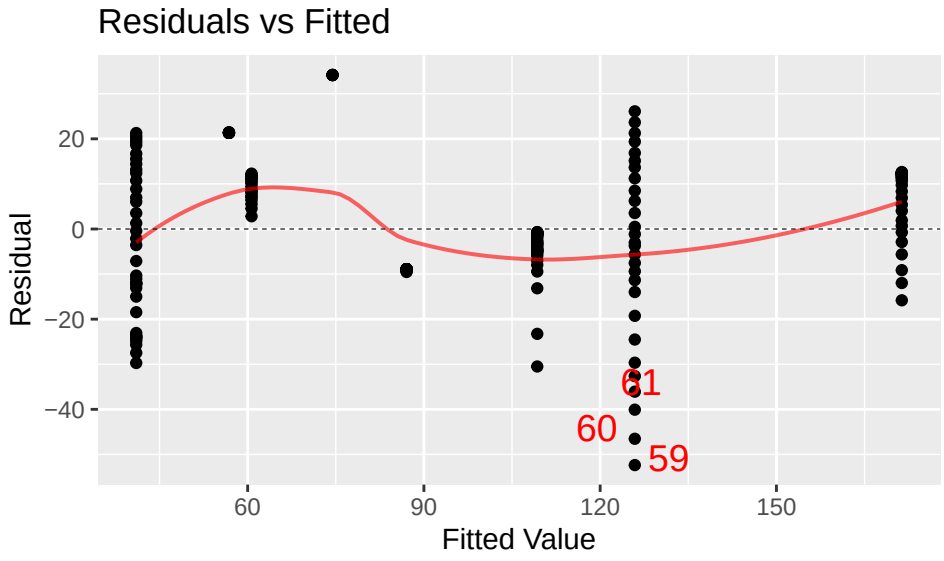



```
favstats(sqrt_fishcount ~ species, data = williamettedata)
```

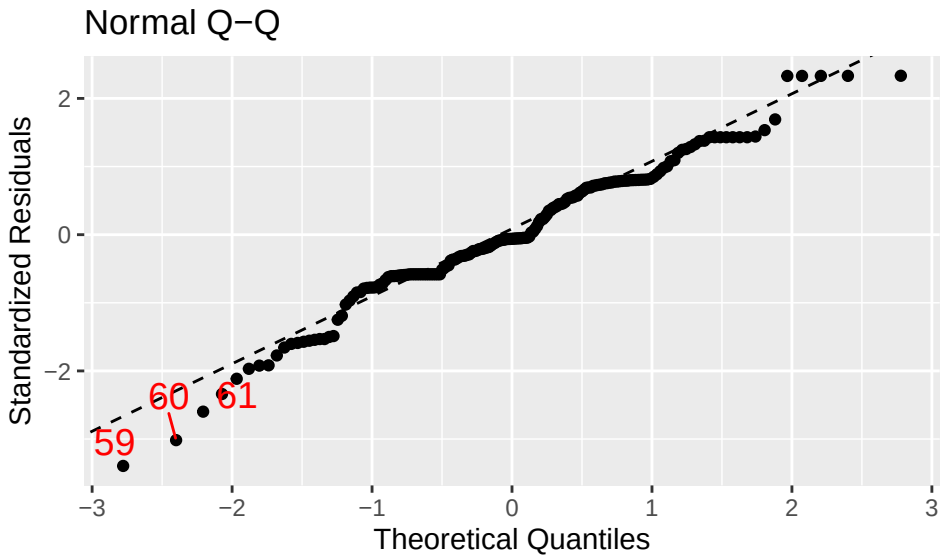
```
##          species      min        Q1    median        Q3       max      mean
## 1           Coho 78.79721 104.48564 107.10002 108.41356 108.62320 105.14239
## 2   Fall Chinook 77.51774  78.14730  78.14730  78.14730  78.30070  78.09477
## 3 Spring Chinook 73.53231 122.69948 153.75917 182.13653 183.97826 148.61843
## 4      Steelhead 11.31371  28.94821  52.60061  67.56057  72.95204  47.46752
##           sd  n missing
## 1  5.8835141 42       0
## 2  0.1521659 27       0
## 3 33.4806528 56       0
## 4 20.7683100 58       0
```

```
mplot(anova2, which = 1)
```

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




```
mplot(anova2, which = 2)
```



Randomization-Based Procedure for MLR For the randomization test for F, we shuffled the response variable so as to keep the relationships between the 4 predictor variables intact. The randomization procedure shows us what we would expect if there was no relationship between the predictor variables and the response variable. Because our F-value is in the extreme end and with an emprical value of 0, we can reject the null model and conclude that at least one of the predictors in our MLR model is useful for predicting square root of fish count.

Note that the t-test for slope for each individual predictor in the model gave enough evidence to suggest the usefulness of all of the predictors in the presence of the other predictors in the model.

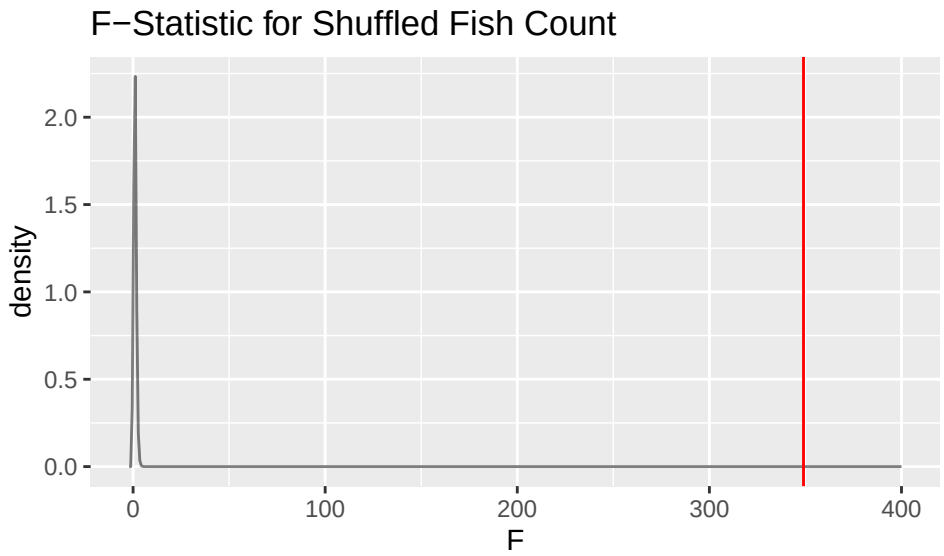
```
set.seed(160)
slopetest <- do(10000)*(lm(shuffle(sqrt_fishcount) ~ time_of_year + log_flow + river_temperature + spec
                           data = williamettedata))
names(slopetest)
```

```
##  [1] "Intercept"             "time_of_yearSpring"    "time_of_yearSummer"
##  [4] "time_of_yearWinter"    "log_flow"              "river_temperature"
```

```
##  [7] "speciesSpring.Chinook" "speciesSteelhead"      "sigma"
## [10] "r.squared"             "F"                     "numdf"
## [13] "dendf"                 ".row"                  ".index"
```

```
gf_dens(~ F, data = slopetest) %>%
  gf_labs(title = "F-Statistic for Shuffled Fish Count") %>%
  gf_lims(x = c(-2, 400)) %>%
  gf_vline(xintercept = 349, color = "red")
```

```
## Warning: geom_vline(): Ignoring `mapping` because `xintercept` was provided.
```




```
pdata(~ F, 349, data = slopetest, lower.tail = FALSE)
```

```
## [1] 0
```

Randomization-Based Procedure for ANOVA Lastly, we conduct a randomization F-test for our two-way additive ANOVA model. We check the F-test for each of our 2 predictor variables: `species` and `time of year`. For the randomization test for F, we shuffled the response variable. The randomization procedure shows us what we would expect if there was no relationship between the predictor variable and the response variable.

For both of the categorical predictors, the respective empirical p-values are 0 as the respective F-statistics are in the extreme end. For both of the categorical predictors, there is sufficient evidence to reject the null hypothesis and support the alternative hypothesis that there is at least one significant difference in means for each of the predictors. Tukey's HSD in the ANOVA section above detected those differences.

```
res <- anova(anova2)$"F value"
res
```

```
## [1] 411.3992  60.0291       NA
```

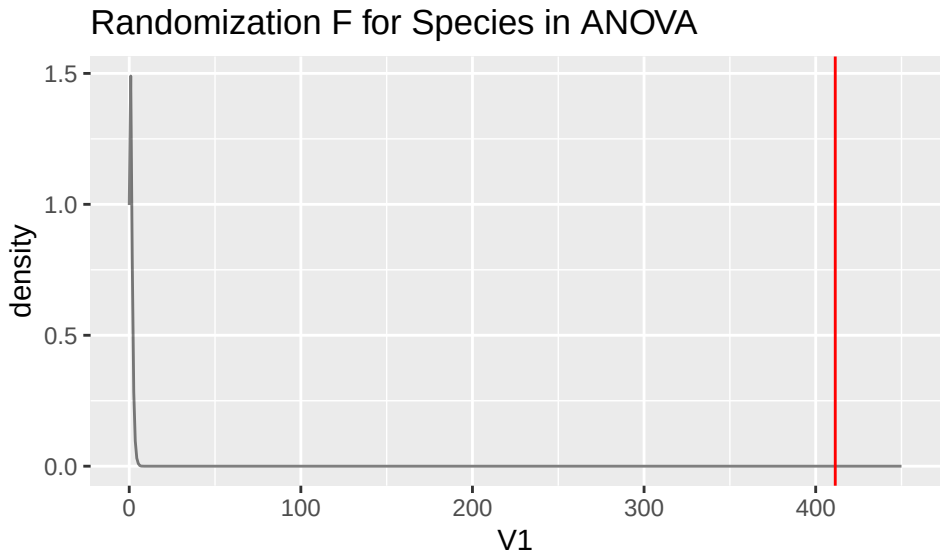
```
newanova <- lm(shuffle(sqrt_fishcount) ~ species + time_of_year, data = williamettedata)
res2 <- anova(newanova)$"F value"
res2
```

```
## [1] 0.3107454 0.0108019        NA
```

```
set.seed(40)
t <- do(10000) * (anova(lm(shuffle(sqrt_fishcount) ~ species + time_of_year, data = williamettedata))$"
```

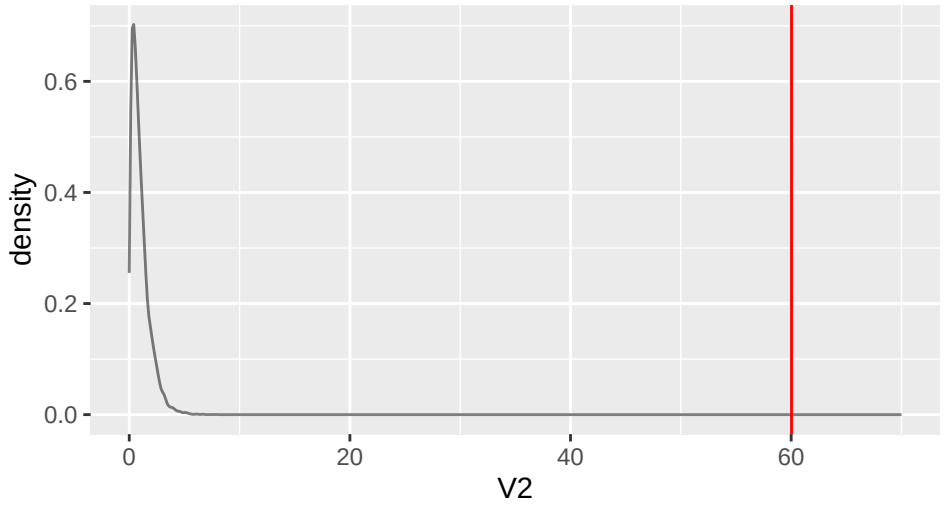
```r
# Species
t <- as.data.frame(t)
gf_dens(~ V1, data = t) %>%
gf_lims(x = c(0, 450)) %>%
gf_labs(title = "Randomization F for Species in ANOVA") %>%
gf_vline(xintercept = 411.40, color = "red")
```

```
## Warning: geom_vline(): Ignoring `mapping` because `xintercept` was provided.
```




```r
# Time of Year
t <- as.data.frame(t)
gf_dens(~ V2, data = t) %>%
gf_lims(x = c(0, 70)) %>%
gf_labs(title = "Randomization F for Time of Year in ANOVA") %>%
gf_vline(xintercept = 60.03, color = "red")
```

```
## Warning: geom_vline(): Ignoring `mapping` because `xintercept` was provided.
```



```
# Species
pdata(~ V1, res[1], data = t, lower.tail = FALSE)
```

```
## [1] 0
```

```
# Time of Year
pdata(~ V2, res[2], data = t, lower.tail = FALSE)
```

```
## [1] 0
```

Conclusion Our group examined what factors affected Pacific Northwest fish counts of various salmon* species. Through use of an MLR model, we found that the following factors were helpful in predicting our response variable, the square root of the fish count, at an alpha level of .05:

The time of year being spring, summer, or winter.

The log of the river flow volume.

The river temperature.

The species being Spring Chinook or Steelhead.

Through the use of a two-way ANOVA with the categorical variables used in our MLR model, we confirmed that at least one mean square root fish count was different for the various fish species as well as the times of year. Tukey's HSD tells us that all comparisons between species means are significant, and that almost all comparisons between time of year means are significant: only spring versus fall and winter versus spring did not show statistical significance. For both our MLR and ANOVA models, we reject the null hypothesis. There is evidence that each of the variables used in our final MLR model is useful in predicting square root fish count, and that at least one mean square root fish count is different between fish species and between times of year for our ANOVA model.

We can only apply our findings to Coho, Fall Chinook, Spring Chinook, and Steelhead in the north Willamette River, as those were the species and location with counts that we ended up using in our models. It is also important to remember that we had to re-express some of our variables, and that we are, for example, predicting the square root of fish count rather than fish count itself. Additionally, we were missing some data points due to government restrictions on data availability, and at times our conditions were not automatically met and required randomization procedures to support the validity of our results.

Ultimately, we found that we can expect the square root of the counts of Coho, Fall Chinook, Spring Chinook, and Steelhead salmon to be affected by the time of year, log of the river flow volume, and river temperature, as well as what species they are. We can use these findings to explore possibilities for salmon conservation in the area and inform future studies and statistical models that work to assess or predict salmon counts.

Our models did a very good job of explaining variation in the data, so the next thing we would like to do is amplify their explanatory power through the inclusion of more rivers, species, and fish counts. Additionally, there are probable causes of salmon decline that we were not able to include in our model, such as presence of logging and fishing pressure, that we would be able to include in a model that referenced material from a broader geographic area.

*We are aware that there is some controversy over the potential classification of Steelhead as a salmon species. Their anadromous nature results in them occupying a very similar niche to salmon, and thus necessitating similar conservation measures. Additionally, recent evidence shows that they are more closely related to salmon than trout, so it seems logical to group them with salmon for inference purposes related to conservation.

References United States. Army. Corps of Engineers. Columbia River Basin Water Management. Portland District, US Army Corps of Engineers, http://pweb.crohms.org/tmt/documents/FPOM/2010/Willamette_Coordination/Willamette%20HMT/Fish%20counts/.

U.S. Geological Survey, 2020, USGS Current Conditions for USGS 14207740 Willamette River above Falls, at Oregon City, OR, https://waterdata.usgs.gov/or/nwis/uv/?site_no=14207740&agency_cd=USGS&%3Brefe rred_module=qw.

U.S. Geological Survey, 2020, USGS Current Conditions for USGS 14163150 Mckenzie River Blw Leaburg Dam, Nr Leaburg, OR, https://waterdata.usgs.gov/or/nwis/uv?cb_00010=on&cb_00060=on&cb_00065=o n&cb_00095=on&format=html&site_no=14163150&period=&begin_date=2021-10-09&end_date=2021- 10-16.

"Willamette Falls Fish Counts." Willamette Falls Fish Counts | Oregon Department of Fish & Wildlife, https://myodfw.com/willamette-falls-fish-counts.