STATS 763 Assign 4

zsye658 Zain Raza Raza Syed

2023-10-20

Survival of the fishest

Some simple Data Exploration & Cleaning

For example with colnames, types etc

```
colnames(fish)
```

```
##  [1] "Fish_no"               "Trip_ID"               "Haul_ID"
##  [4] "Temp_water"            "Survival_time..h."     "Status_end..1...dead."
##  [7] "Vitality_class"        "Wave_height..cm."      "Depth..m."
## [10] "Processing.time..min." "Hopper"                "Haul_duration..min."
## [13] "Substrate"             "Length..cm."
```

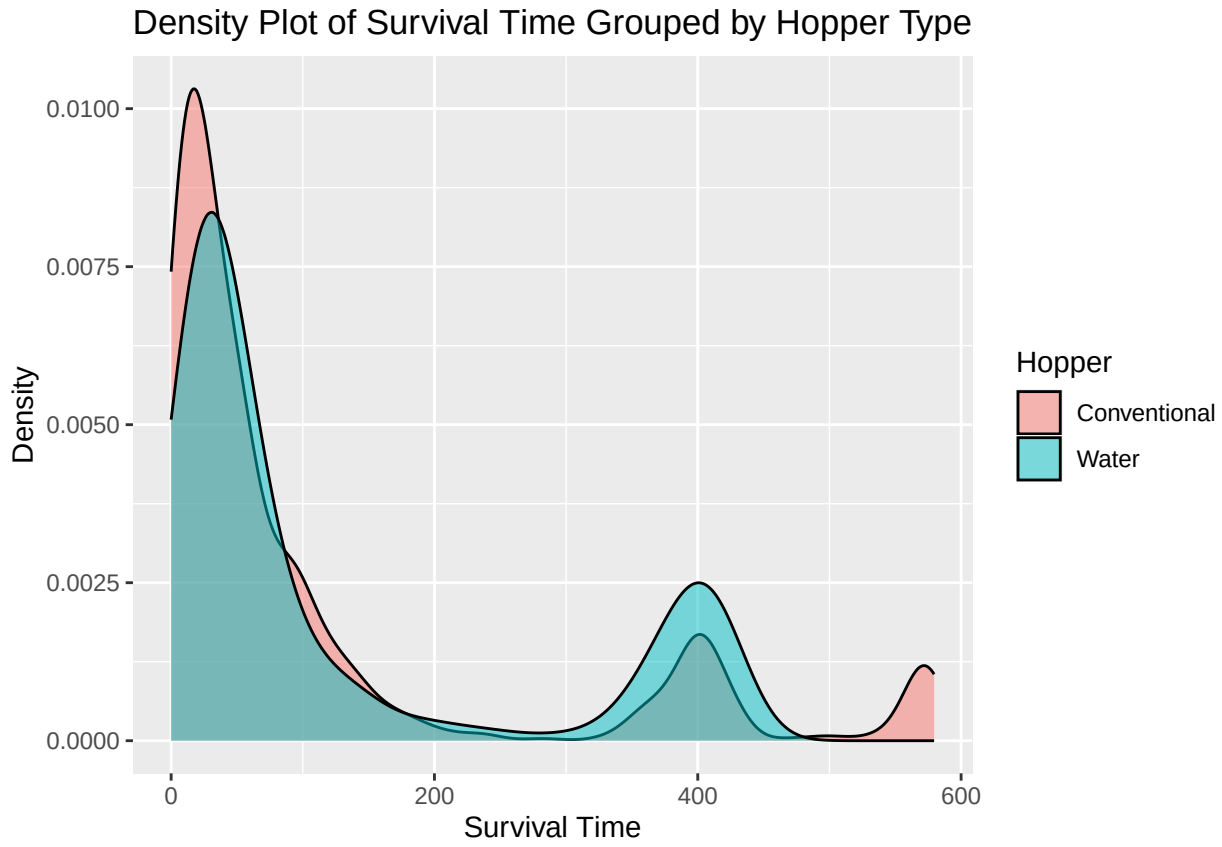
```
str(fish)
```

```
## Classes 'data.table' and 'data.frame':   1292 obs. of  14 variables:
##  $ Fish_no              : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ Trip_ID              : chr  "1ovl1" "1ovl1" "1ovl1" "1ovl1" ...
##  $ Haul_ID              : chr  "1ovl1_7" "1ovl1_7" "1ovl1_7" "1ovl1_7" ...
##  $ Temp_water           : num  14.5 14.5 14.5 14.5 14.5 14.5 14.5 14.5 14.5 14.5 ...
##  $ Survival_time..h.    : num  571.4 571.5 0 51.9 40.7 ...
##  $ Status_end..1...dead.: int  1 1 0 0 0 0 1 0 0 0 ...
##  $ Vitality_class       : chr  "C" "A" "C" "C" ...
##  $ Wave_height..cm.     : int  91 91 91 91 91 91 91 91 91 91 ...
##  $ Depth..m.            : int  33 33 33 33 33 33 33 33 33 33 ...
##  $ Processing.time..min.: int  14 14 14 7 14 14 7 7 7 14 ...
##  $ Hopper               : chr  "Conventional" "Conventional" "Conventional" "Conventional" ...
##  $ Haul_duration..min.  : int  105 105 105 105 105 105 105 105 105 105 ...
##  $ Substrate            : chr  "Sand" "Sand" "Sand" "Sand" ...
##  $ Length..cm.          : int  21 22 26 20 25 20 25 18 20 22 ...
##  - attr(*, ".internal.selfref")=<externalptr>
```

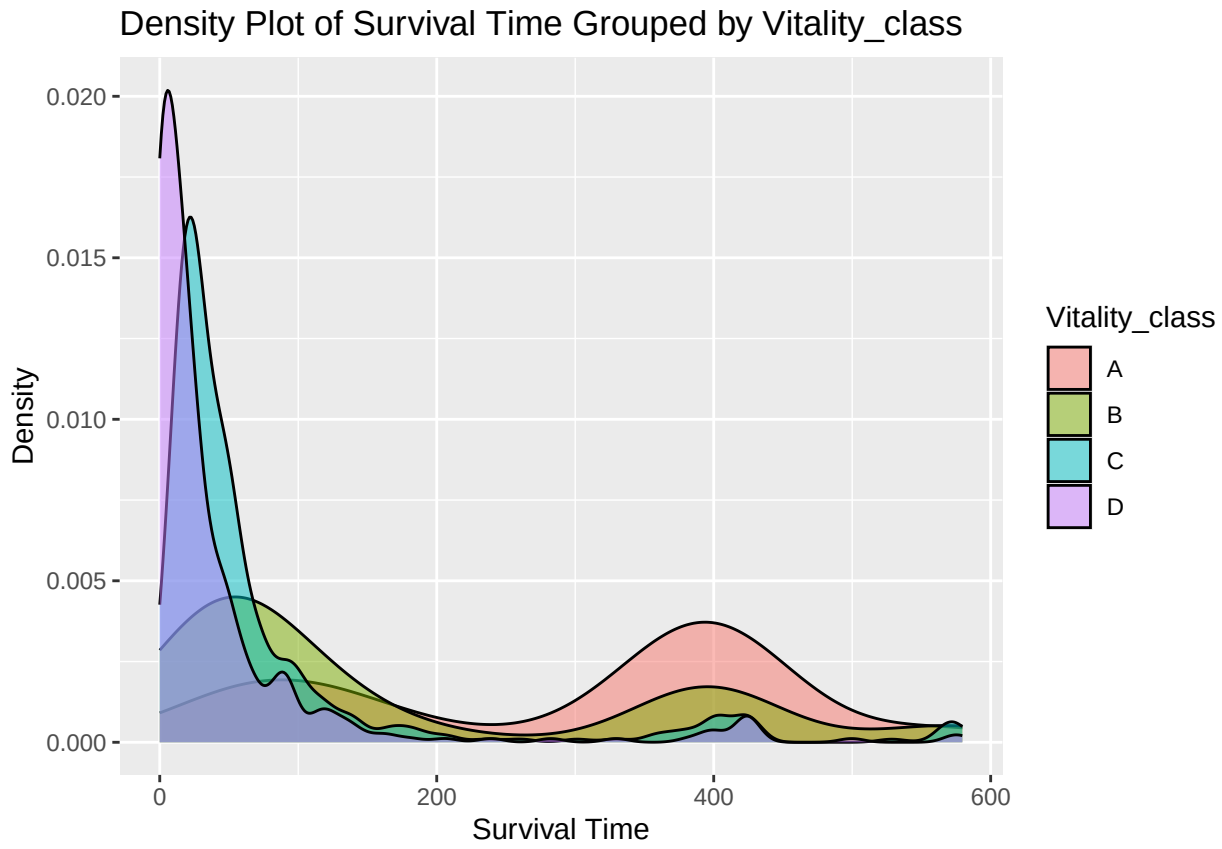
```
#changing vitality class
colnames(fish)[6] = "dead"
fish$dead = 1 * !fish$dead
colnames(fish)[5] = "Survival_Time"
colnames(fish)[8] = "Wave_height"
colnames(fish)[9] = "Depth"
colnames(fish)[10] = "Processing_time"
colnames(fish)[12] = "Haul_duration"
colnames(fish)[14] = "Length"
```

Data Exploration and visualisation

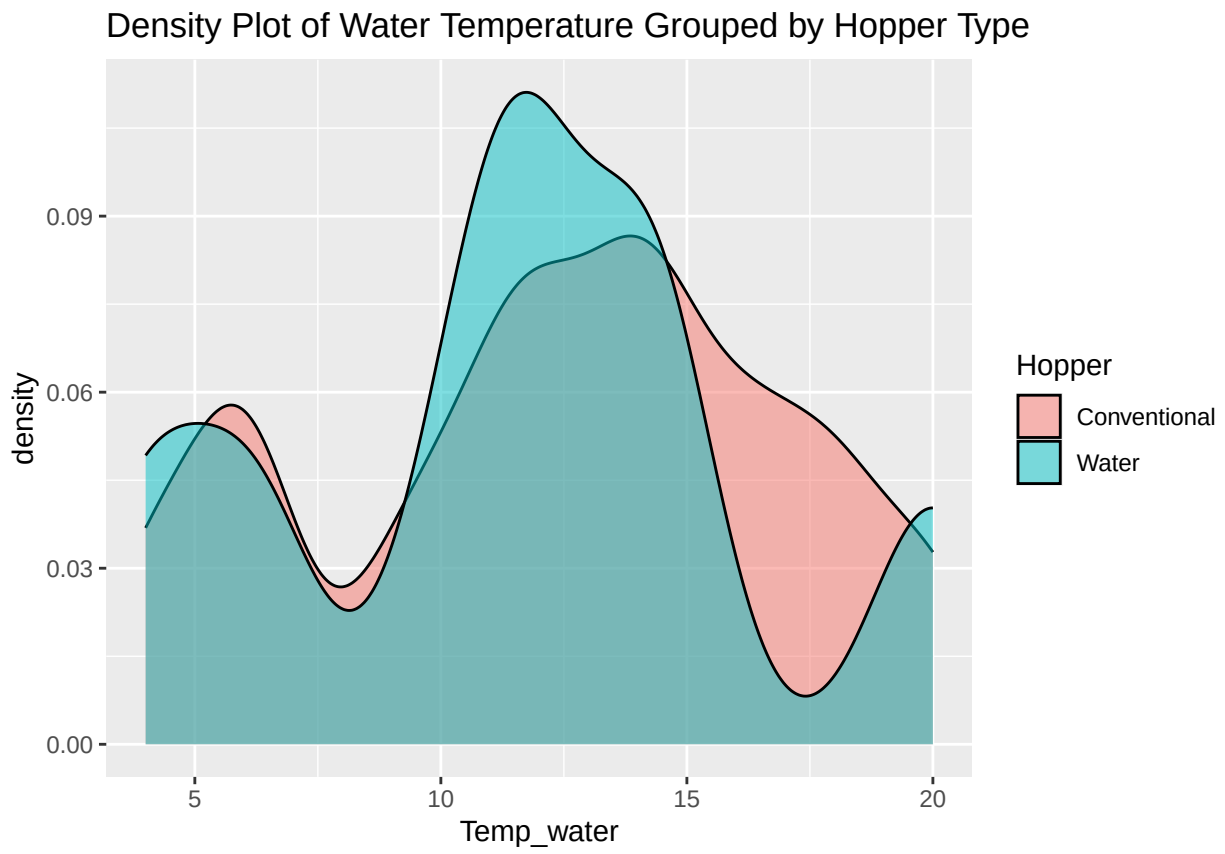
```
# Some meaningless graphs
ggplot(fish, aes(x = Survival_Time, fill = Hopper)) +
  geom_density(alpha = 0.5) +
  labs(title = "Density Plot of Survival Time Grouped by Hopper Type",
       x = "Survival Time",
       y = "Density")
```



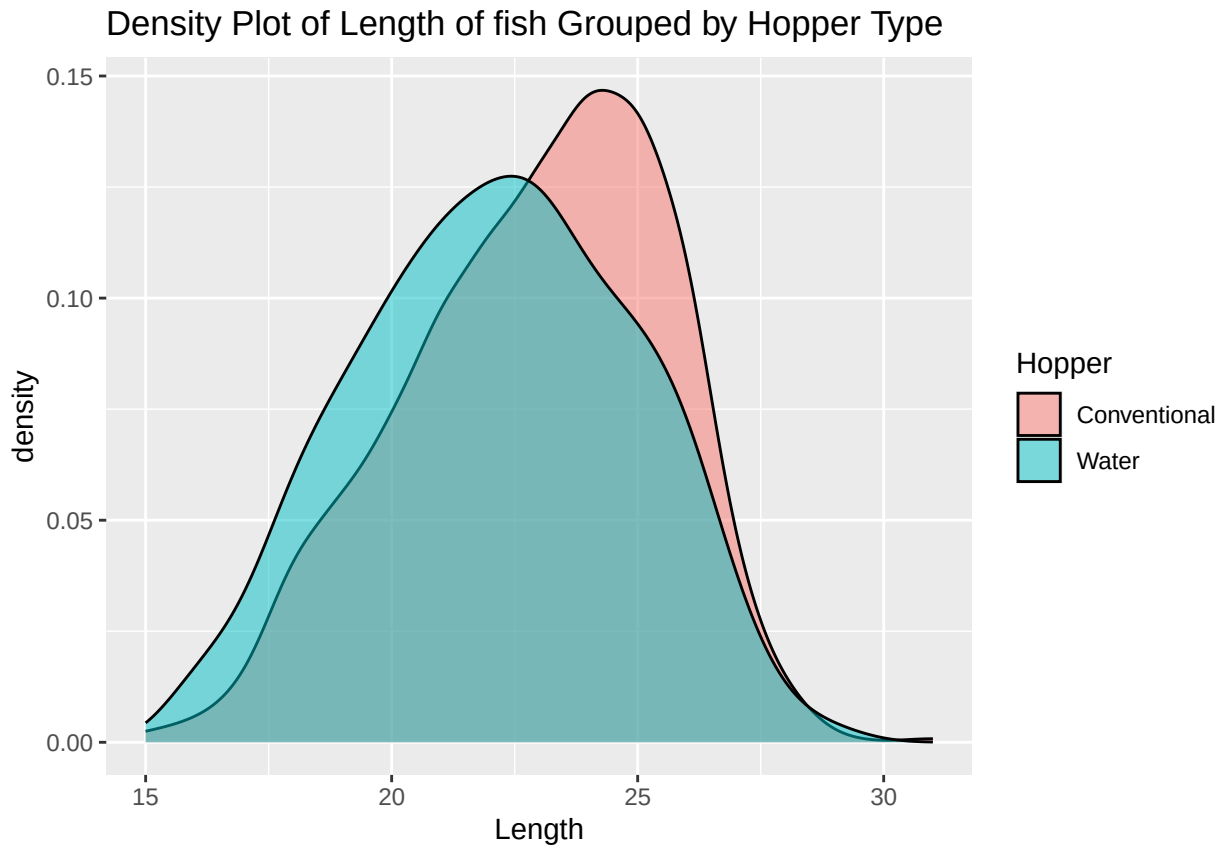

```
ggplot(fish, aes(x = Survival_Time, fill = Vitality_class)) +
  geom_density(alpha = 0.5) +
  labs(title = "Density Plot of Survival Time Grouped by Vitality_class",
       x = "Survival Time",
       y = "Density")
```



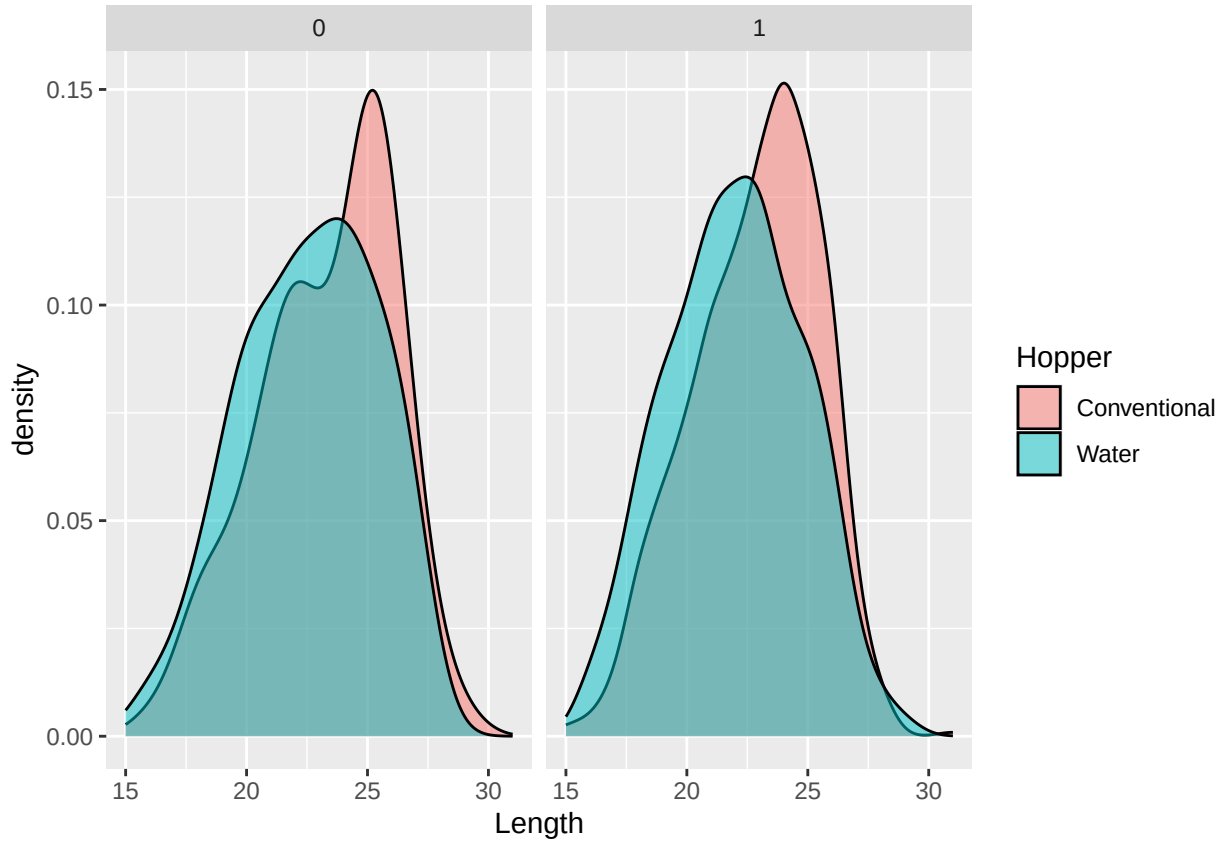

```
ggplot(fish, aes(x = Temp_water, fill = Hopper )) +
  geom_density(alpha=0.5) +
  labs(title = "Density Plot of Water Temperature Grouped by Hopper Type")
```




```
ggplot(fish, aes(x = Length, fill = Hopper))  + geom_density(alpha=0.5) +
  ggtitle("Density Plot of Length of fish Grouped by Hopper Type")
```



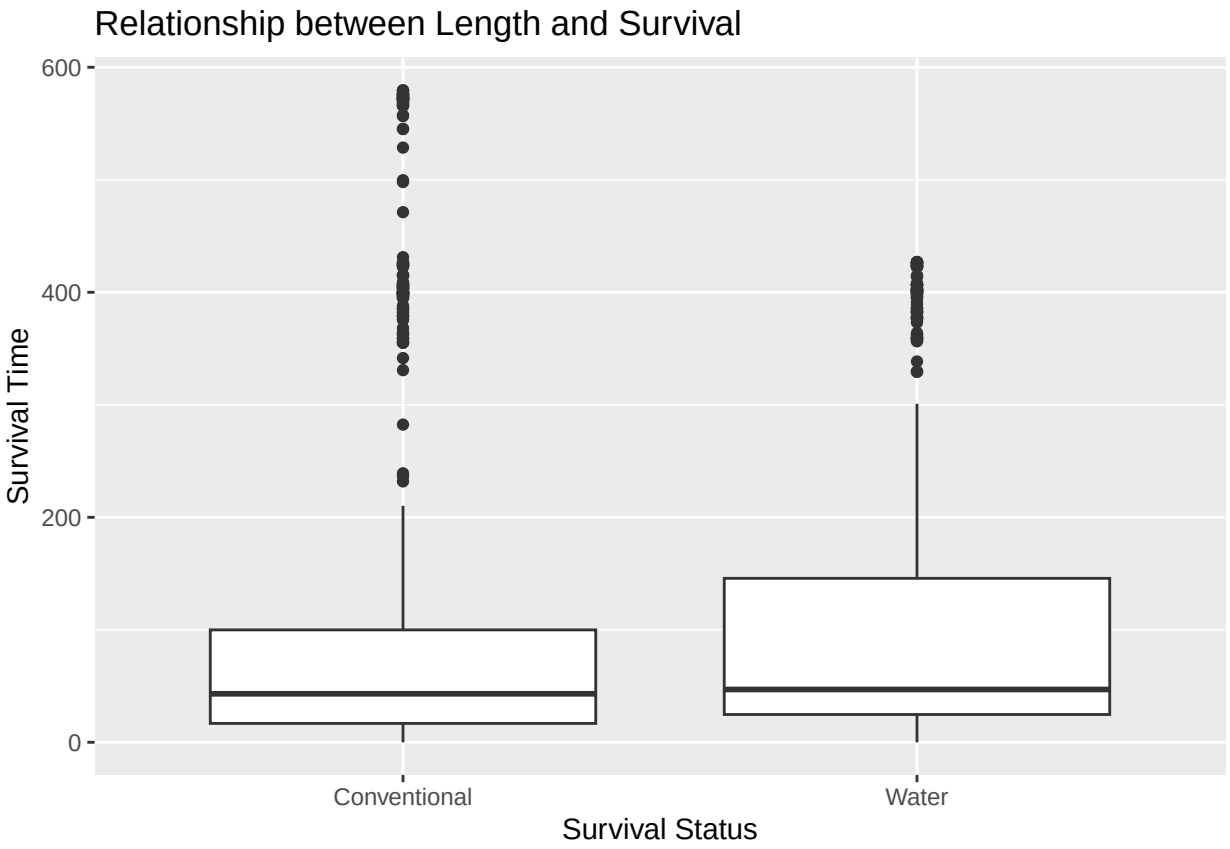

```
ggplot(fish, aes(x = Length, fill = Hopper))  + geom_density(alpha=0.5) + facet_wrap(~dead)
```



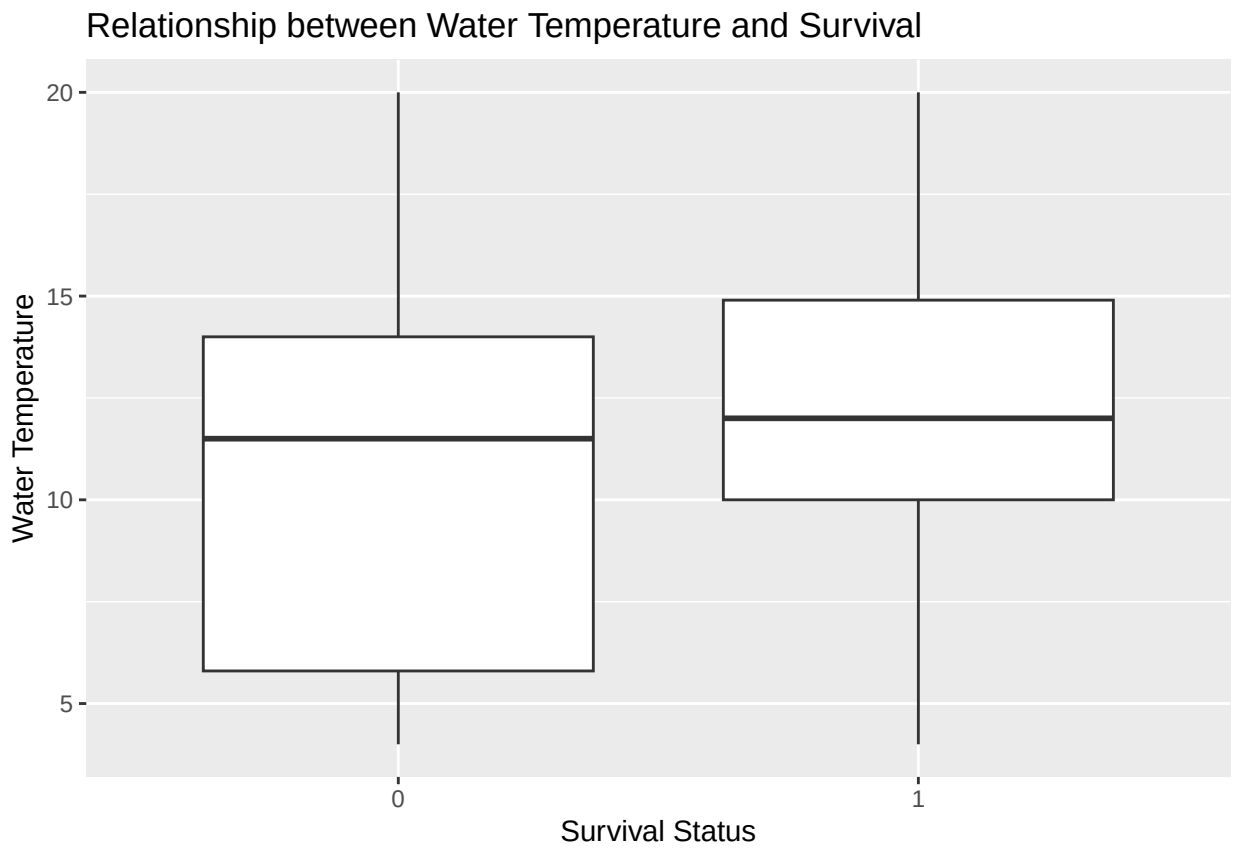

```
  ggtitle("Distribution is same for dead and not dead fish")
```

```
## $title
## [1] "Distribution is same for dead and not dead fish"
##
## attr(,"class")
## [1] "labels"
```

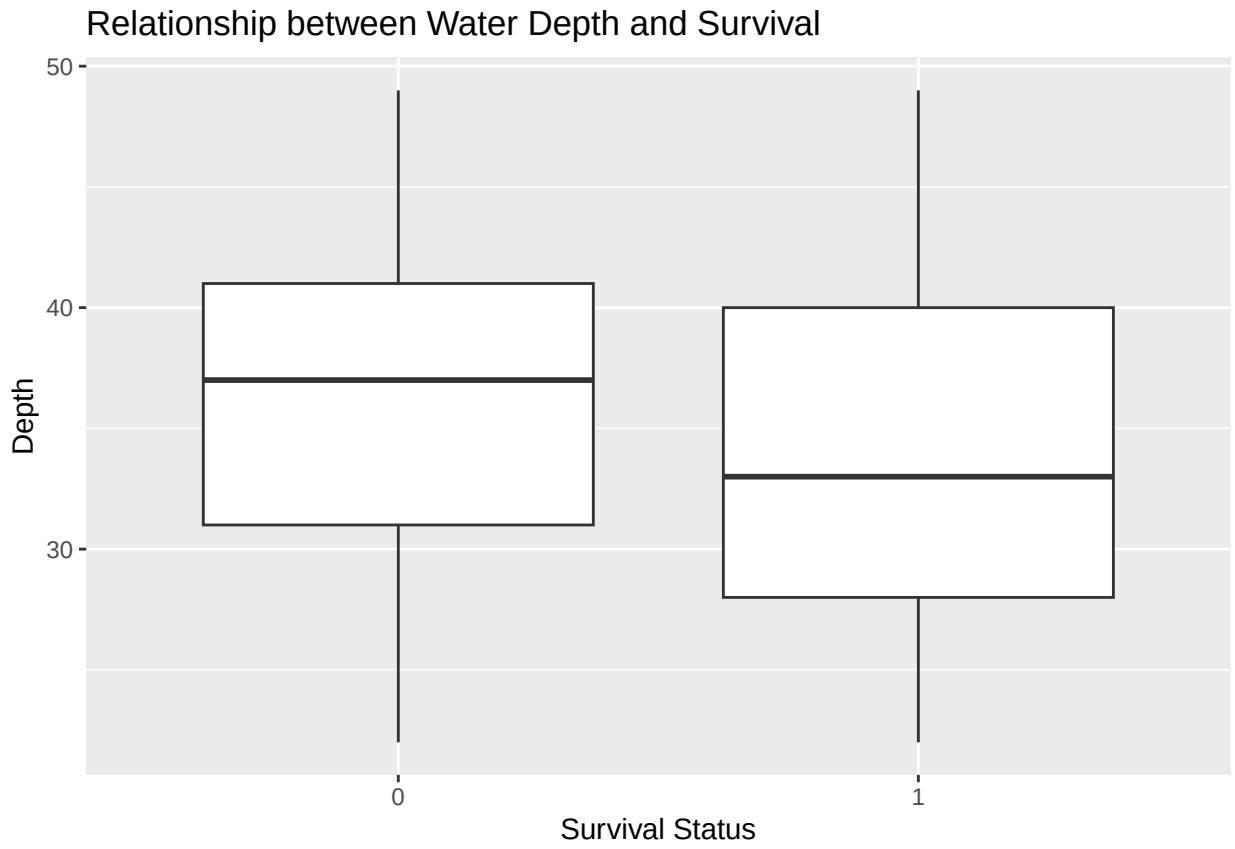
```
# Boxplot for Length against Survival
ggplot(fish, aes(x = Hopper, y = Survival_Time)) +
  geom_boxplot() +
  xlab("Survival Status") +
  ylab("Survival Time") +
  ggtitle("Relationship between Length and Survival")
```




```
# Boxplot for Temp_water against Survival
ggplot(fish, aes(x = as.factor(`dead`), y = Temp_water)) +
  geom_boxplot() +
  xlab("Survival Status") +
  ylab("Water Temperature") +
  ggtitle("Relationship between Water Temperature and Survival")
```




```
# ... similarly for other variables
ggplot(fish, aes(x = as.factor(`dead`), y = Depth)) +
  geom_boxplot() +
  xlab("Survival Status") +
  ylab("Depth") +
  ggtitle("Relationship between Water Depth and Survival")
```



1. Which variables are potential confounders for the effect of hopper type on survival?

Confounders are variables that are related to both the treatment (hopper type in this case) and the outcome (survival), but are not in the causal pathway from the treatment to the outcome. Potential confounders could be:

Temp_water: The water temperature might affect both the choice of hopper type and the survival of the fish. Wave_height: The height of waves at sea could influence both the hopper type used and fish survival. Depth: The depth at which fish are caught could relate to both hopper type and survival. Substrate: The type of sea bottom where fish are caught might affect both hopper type and fish survival. Haul_duration: The duration of the trawling operation could affect both the hopper type and survival.

2. Which variables are potential mediators of the effect of hopper type on survival?

Mediators are variables that are affected by the treatment and, in turn, affect the outcome. Potential mediators could be:

Vitality_class: The health of the fish after being caught might be affected by the hopper type and could subsequently influence their survival. Processing time: The time from when the fish come on board to when they end up in storage tanks might depend on the hopper type and could affect survival.

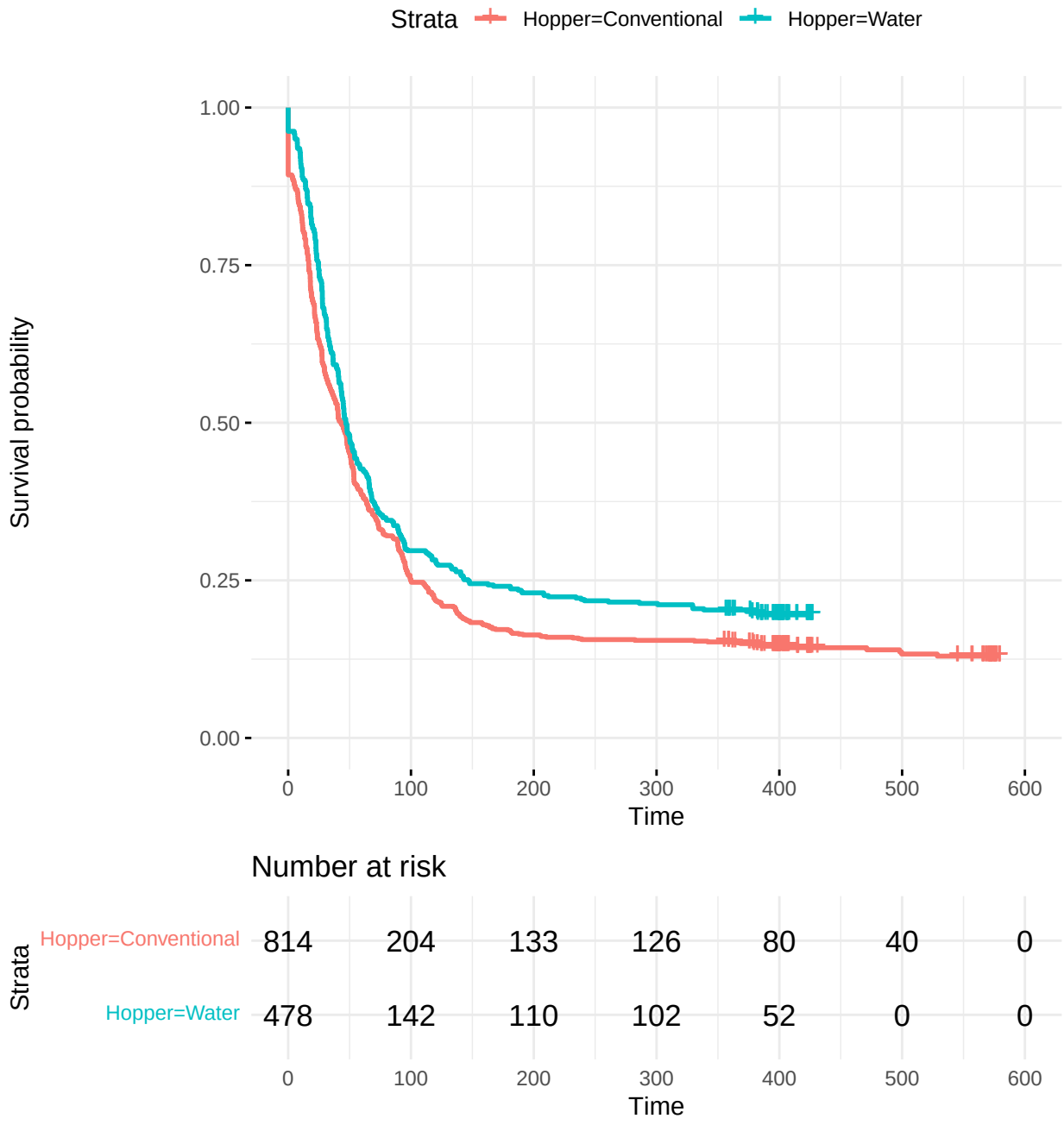
3. Estimate the effect of hopper type on survival.

```r
mo1 = coxph(Surv(Survival_Time, dead) ~ Hopper, fish)
summary(mo1)
```

```
## Call:
## coxph(formula = Surv(Survival_Time, dead) ~ Hopper, data = fish)
##
##   n= 1292, number of events= 1084
##
##                 coef exp(coef) se(coef)      z Pr(>|z|)
## HopperWater -0.18420   0.83177  0.06361 -2.896  0.00378 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##             exp(coef) exp(-coef) lower .95 upper .95
## HopperWater    0.8318      1.202    0.7343    0.9422
##
## Concordance= 0.528  (se = 0.008 )
## Likelihood ratio test= 8.52  on 1 df,   p=0.004
## Wald test            = 8.39  on 1 df,   p=0.004
## Score (logrank) test = 8.41  on 1 df,   p=0.004
```

```r
# Fit the Kaplan-Meier survival curves for Hopper Type
km_fit_hopper <- survfit(Surv(Survival_Time, dead) ~ Hopper, data = fish)

# Plot the survival curves for Hopper Type
ggsurvplot(km_fit_hopper, data = fish, risk.table = TRUE, ggtheme = theme_minimal())
```




```
big_mo = coxph(Surv(Survival_Time, dead) ~ Hopper + Length + Temp_water + Wave_height + Depth + Substra
summary(big_mo)
```

```
## Call:
## coxph(formula = Surv(Survival_Time, dead) ~ Hopper + Length +
##     Temp_water + Wave_height + Depth + Substrate + Haul_duration +
##     Vitality_class + Processing_time, data = fish)
##
##   n= 1292, number of events= 1084
##
##                                   coef  exp(coef)   se(coef)      z Pr(>|z|)
```

```
## HopperWater                  -1.022e-01  9.028e-01  6.522e-02 -1.568   0.1169
## Length                       -7.695e-02  9.259e-01  1.157e-02 -6.649 2.95e-11
## Temp_water                    1.260e-01  1.134e+00  8.112e-03 15.537  < 2e-16
## Wave_height                   8.538e-05  1.000e+00  6.851e-04  0.125   0.9008
## Depth                        -8.938e-03  9.911e-01  4.938e-03 -1.810   0.0703
## SubstrateMud to \nmuddy Sand -3.166e-01  7.286e-01  2.214e-01 -1.430   0.1526
## SubstrateSand                -2.743e-02  9.729e-01  1.030e-01 -0.266   0.7899
## Haul_duration                 3.091e-02  1.031e+00  6.089e-03  5.076 3.85e-07
## Vitality_classB               1.047e+00  2.850e+00  1.585e-01  6.609 3.87e-11
## Vitality_classC               2.137e+00  8.473e+00  1.580e-01 13.524  < 2e-16
## Vitality_classD               3.115e+00  2.253e+01  1.631e-01 19.093  < 2e-16
## Processing_time               1.051e-02  1.011e+00  4.761e-03  2.208   0.0273
##
## HopperWater
## Length                       ***
## Temp_water                   ***
## Wave_height
## Depth                        .
## SubstrateMud to \nmuddy Sand
## SubstrateSand
## Haul_duration                ***
## Vitality_classB              ***
## Vitality_classC              ***
## Vitality_classD              ***
## Processing_time              *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##                              exp(coef) exp(-coef) lower .95 upper .95
## HopperWater                     0.9028    1.10766    0.7945    1.0259
## Length                          0.9259    1.07999    0.9052    0.9472
## Temp_water                      1.1343    0.88158    1.1164    1.1525
## Wave_height                     1.0001    0.99991    0.9987    1.0014
## Depth                           0.9911    1.00898    0.9816    1.0007
## SubstrateMud to \nmuddy Sand    0.7286    1.37252    0.4721    1.1243
## SubstrateSand                   0.9729    1.02781    0.7951    1.1906
## Haul_duration                   1.0314    0.96957    1.0192    1.0438
## Vitality_classB                 2.8501    0.35087    2.0891    3.8882
## Vitality_classC                 8.4726    0.11803    6.2162   11.5480
## Vitality_classD                22.5349    0.04438   16.3675   31.0262
## Processing_time                 1.0106    0.98954    1.0012    1.0200
##
## Concordance= 0.785  (se = 0.006 )
## Likelihood ratio test= 902.7  on 12 df,   p=<2e-16
## Wald test            = 834.1  on 12 df,   p=<2e-16
## Score (logrank) test = 873.4  on 12 df,   p=<2e-16
```

The negative coefficient suggests that the water-filled hopper is associated with a decrease in the hazard, indicating better survival compared to the reference category (likely a dry hopper)

However, in the extended model with additional covariates (such as Length, Temp_water, Wave_height, etc.), the coefficient for Hopper changed to -0.1022 and was not statistically significant (p=0.1169). This implies that once we control for these additional factors, the type of hopper may not independently impact fish survival.

4. How does the effect of hopper type on survival vary with fish characteristics and sea conditions?

To answer this, interaction terms between hopper type and other variables (e.g., Temp_water, Wave_height, Depth, Length) could be added to the model.

```
# Fit the Cox model with interaction between Hopper and Length
mo1_length <- coxph(Surv(Survival_Time, dead) ~ Hopper * Length, data = fish)
summary(mo1_length)
```

```
## Call:
## coxph(formula = Surv(Survival_Time, dead) ~ Hopper * Length,
##     data = fish)
##
##   n= 1292, number of events= 1084
##
##                        coef exp(coef) se(coef)      z Pr(>|z|)
## HopperWater         0.30549   1.35729  0.51398  0.594   0.5523
## Length             -0.02832   0.97208  0.01375 -2.059   0.0395 *
## HopperWater:Length -0.02303   0.97723  0.02286 -1.008   0.3136
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##                    exp(coef) exp(-coef) lower .95 upper .95
## HopperWater           1.3573     0.7368    0.4956    3.7169
## Length                0.9721     1.0287    0.9462    0.9986
## HopperWater:Length    0.9772     1.0233    0.9344    1.0220
##
## Concordance= 0.547  (se = 0.009 )
## Likelihood ratio test= 20.59  on 3 df,   p=1e-04
## Wald test            = 19.86  on 3 df,   p=2e-04
## Score (logrank) test = 20  on 3 df,   p=2e-04
```

```
# Fit the Cox model with interaction between Hopper and Temp_water
mo1_temp <- coxph(Surv(Survival_Time, dead) ~ Hopper * Temp_water, data = fish)
summary(mo1_temp)
```

```
## Call:
## coxph(formula = Surv(Survival_Time, dead) ~ Hopper * Temp_water,
##     data = fish)
##
##   n= 1292, number of events= 1084
##
##                              coef exp(coef)  se(coef)      z Pr(>|z|)
## HopperWater             0.084451  1.088119  0.175264  0.482    0.630
## Temp_water              0.080519  1.083849  0.008325  9.672   <2e-16 ***
## HopperWater:Temp_water -0.020589  0.979622  0.013429 -1.533    0.125
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##                        exp(coef) exp(-coef) lower .95 upper .95
## HopperWater               1.0881     0.9190    0.7718     1.534
## Temp_water                1.0838     0.9226    1.0663     1.102
## HopperWater:Temp_water    0.9796     1.0208    0.9542     1.006
##
## Concordance= 0.643  (se = 0.009 )
```

```
## Likelihood ratio test= 135.3  on 3 df,   p=<2e-16
## Wald test            = 135.9  on 3 df,   p=<2e-16
## Score (logrank) test = 139  on 3 df,   p=<2e-16
```

```
# Fit the Cox model with interaction between Hopper and Haul_duration
mo1_haul <- coxph(Surv(Survival_Time, dead) ~ Hopper * Haul_duration, data = fish)
summary(mo1_haul)
```

```
## Call:
## coxph(formula = Surv(Survival_Time, dead) ~ Hopper * Haul_duration,
##     data = fish)
##
##   n= 1292, number of events= 1084
##
##                                coef exp(coef)  se(coef)      z Pr(>|z|)
## HopperWater                0.023248  1.023520  1.227756  0.019    0.985
## Haul_duration              0.024312  1.024610  0.005574  4.361 1.29e-05 ***
## HopperWater:Haul_duration -0.002081  0.997922  0.010175 -0.204    0.838
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##                           exp(coef) exp(-coef) lower .95 upper .95
## HopperWater                  1.0235      0.977   0.09226    11.354
## Haul_duration                1.0246      0.976   1.01348     1.036
## HopperWater:Haul_duration    0.9979      1.002   0.97822     1.018
##
## Concordance= 0.556  (se = 0.009 )
## Likelihood ratio test= 34.63  on 3 df,   p=1e-07
## Wald test            = 34.6  on 3 df,   p=1e-07
## Score (logrank) test = 34.71  on 3 df,   p=1e-07
```

```
# Fit the Cox model with all interaction terms
mo1_all <- coxph(Surv(Survival_Time, dead) ~ Hopper * (Length + Temp_water + Haul_duration), data = fis
summary(mo1_all)
```

```
## Call:
## coxph(formula = Surv(Survival_Time, dead) ~ Hopper * (Length +
##     Temp_water + Haul_duration), data = fish)
##
##   n= 1292, number of events= 1084
##
##                                coef exp(coef)  se(coef)      z Pr(>|z|)
## HopperWater                1.702559  5.487974  1.484195  1.147 0.251329
## Length                    -0.017967  0.982194  0.014075 -1.276 0.201781
## Temp_water                 0.078389  1.081543  0.008394  9.339  < 2e-16 ***
## Haul_duration              0.019769  1.019965  0.005745  3.441 0.000579 ***
## HopperWater:Length        -0.038090  0.962626  0.023339 -1.632 0.102671
## HopperWater:Temp_water    -0.016089  0.984040  0.013935 -1.155 0.248284
## HopperWater:Haul_duration -0.007253  0.992773  0.011094 -0.654 0.513224
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##                           exp(coef) exp(-coef) lower .95 upper .95
## HopperWater                  5.4880     0.1822    0.2993   100.638
## Length                       0.9822     1.0181    0.9555     1.010
## Temp_water                   1.0815     0.9246    1.0639     1.099
```

```
## Haul_duration                1.0200     0.9804    1.0085     1.032
## HopperWater:Length           0.9626     1.0388    0.9196     1.008
## HopperWater:Temp_water       0.9840     1.0162    0.9575     1.011
## HopperWater:Haul_duration    0.9928     1.0073    0.9714     1.015
##
## Concordance= 0.648  (se = 0.009 )
## Likelihood ratio test= 162.1  on 7 df,   p=<2e-16
## Wald test            = 156.3  on 7 df,   p=<2e-16
## Score (logrank) test = 160.8  on 7 df,   p=<2e-16
```

Based on these models, it appears that the effect of hopper type on fish survival does not significantly vary with fish length, water temperature, or haul duration. This means that the impact of the hopper type is consistent across these different conditions and characteristics.

5. Is vitality class a good summary of the survival risk of the fish?

```
mo_hopper_vitality = coxph(Surv(Survival_Time, dead) ~ Hopper + Vitality_class, fish)
summary(mo_hopper_vitality)
```

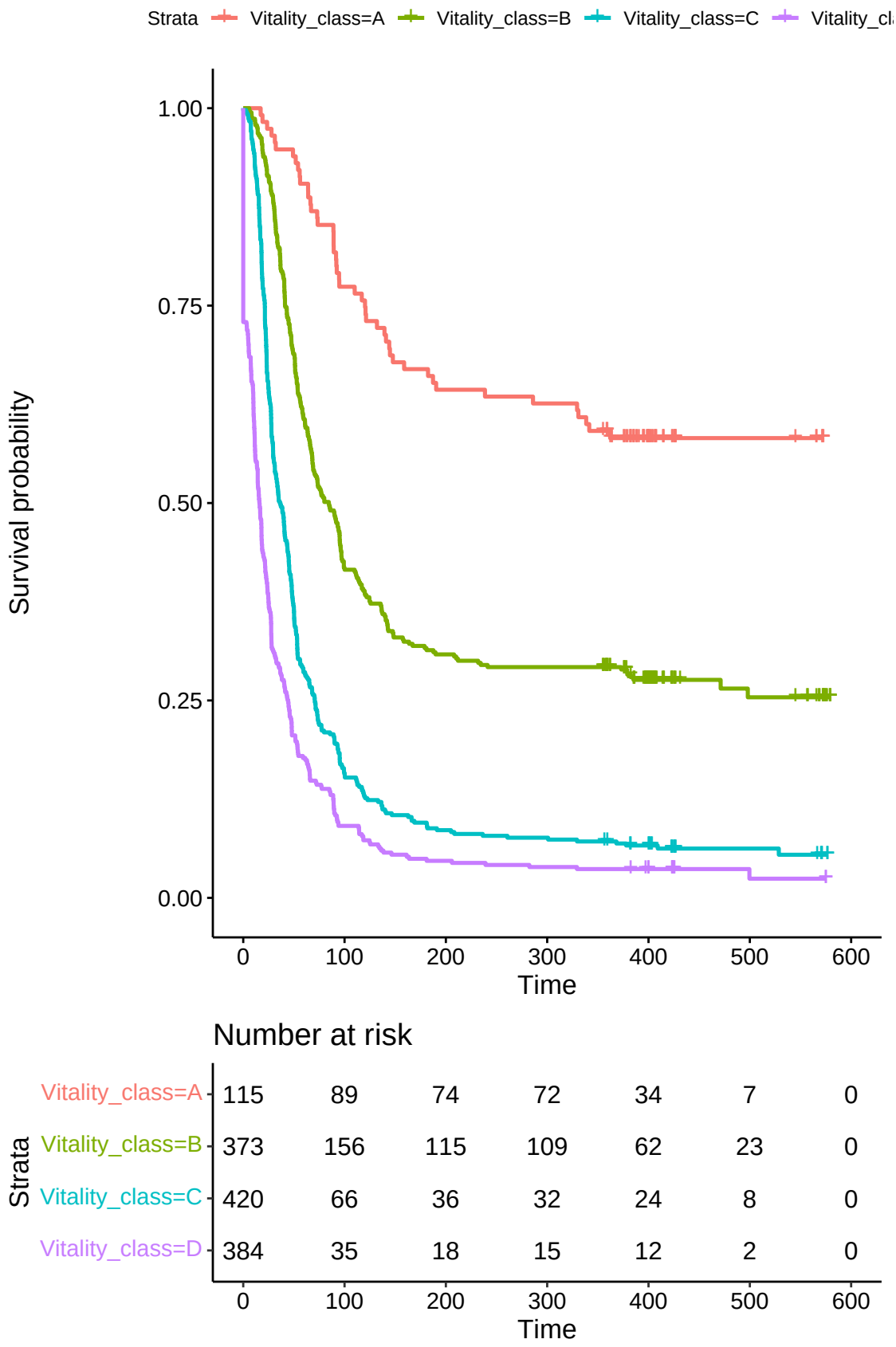
```
## Call:
## coxph(formula = Surv(Survival_Time, dead) ~ Hopper + Vitality_class,
##     data = fish)
##
##   n= 1292, number of events= 1084
##
##                      coef exp(coef) se(coef)      z Pr(>|z|)
## HopperWater     -0.05509   0.94640  0.06394 -0.862    0.389
## Vitality_classB  0.88583   2.42501  0.15728  5.632 1.78e-08 ***
## Vitality_classC  1.67090   5.31697  0.15465 10.804  < 2e-16 ***
## Vitality_classD  2.19123   8.94618  0.15574 14.070  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##                 exp(coef) exp(-coef) lower .95 upper .95
## HopperWater        0.9464     1.0566    0.8349     1.073
## Vitality_classB    2.4250     0.4124    1.7817     3.301
## Vitality_classC    5.3170     0.1881    3.9267     7.200
## Vitality_classD    8.9462     0.1118    6.5928    12.140
##
## Concordance= 0.708  (se = 0.008 )
## Likelihood ratio test= 429.7  on 4 df,   p=<2e-16
## Wald test            = 377.7  on 4 df,   p=<2e-16
## Score (logrank) test = 441.3  on 4 df,   p=<2e-16
```

Fish with a "Vitality_class" of B are about 2.85 times more at risk than those with class A. Fish with a "Vitality_class" of C are about 8.47 times more at risk than those with class A. Fish with a "Vitality_class" of D are about 22.53 times more at risk than those with class A.

The effect size is substantial, especially moving from class B to D, indicating a clear gradient in survival risk.

```
km_fit <- survfit(Surv(Survival_Time, dead) ~ Vitality_class, data = fish)

ggsurvplot(km_fit, data = fish, risk.table = TRUE)
```


Strata
Vitality_class=A
Vitality_class=B
Vitality_class=C
Vitality_cl
Survival probability
1.00
0.75
0.50
0.25
0.00
0
100
200
300
400
500
600
Time
Number at risk
Strata
Vitality_class=A
Vitality_class=B
Vitality_class=C
Vitality_class=D
115 89 74 72 34 7 0
373 156 115 109 62 23 0
420 66 36 32 24 8 0
384 35 18 15 12 2 0
0
100
200
300
400
500
600
Time