Problem set 6: Worked solutions

Statistics and statistical programming
Northwestern University
MTS 525

Aaron Shaw

November 9, 2020

Contents

Programming challenges

PC1 Download, import, and reshape the data

First I can import the data. My preferred method for working with .xls files is (whenever possible) to open them up and save them as .csv or .tsv files. I've done that here and uploaded a copy of the resulting file to the course website to make this example reproducible. I'll use the tidyverse `read_csv()` command to import the data.

```
library(tidyverse)

raw_df <- read_csv("https://communitydata.science/~ads/teaching/2020/stats/data/week_09/owan03.csv")
raw_df
```

```
## # A tibble: 11 x 4
##       X1    X2    X3    X4
##    <dbl> <dbl> <dbl> <dbl>
##  1    70    49    30    34
##  2    77    60    37    36
##  3    83    63    56    48
##  4    87    67    65    48
##  5    92    70    76    65
##  6    93    74    83    91
##  7   100    77    87    98
##  8   102    80    90   102
##  9   102    89    94    NA
```

```
## 10   103    NA    97    NA
## 11    96    NA    NA    NA
```

Now, let's organize the data. We want this in "long" format where the survival times in weeks for each of the three dosage levels are all arranged in a single column and another column labels the corresponding dosage level. You might note that having variable information captured in the column headers is a pretty common situation, so learning how to do this kind of dataset manipulation is a very useful skill!

First, I'll rename the original column names so that the information about dosage is explicitly labeled.

```
colnames(raw_df) <- c("None", "Low", "Medium", "High")
```

Now, I'll use the very handy `pivot_longer()` function from the `tidyverse` library to do the rearranging.

```
df <- raw_df %>%
  pivot_longer(
    cols = everything(),
    names_to = "dose",
    values_to = "weeks_alive"
  )

df
```

```
## # A tibble: 44 x 2
##    dose   weeks_alive
##    <chr>        <dbl>
##  1 None            70
##  2 Low             49
##  3 Medium          30
##  4 High            34
##  5 None            77
##  6 Low             60
##  7 Medium          37
##  8 High            36
##  9 None            83
## 10 Low             63
## # ... with 34 more rows
```

That's looking good! I should probably convert the "dose" variable into a factor while I'm at it. I'll be super explicit about the levels here just to make sure nothing gets lost along the way...

```
df$dose <- factor(df$dose, levels = c("None", "Low", "Medium", "High"))
```

Last, but not least, the NA values from my original dataframe aren't really doing me any good here, so I can drop them:

```
df <- df[complete.cases(df), ]
```

PC2 Summarize the data

Summary statistics overall:

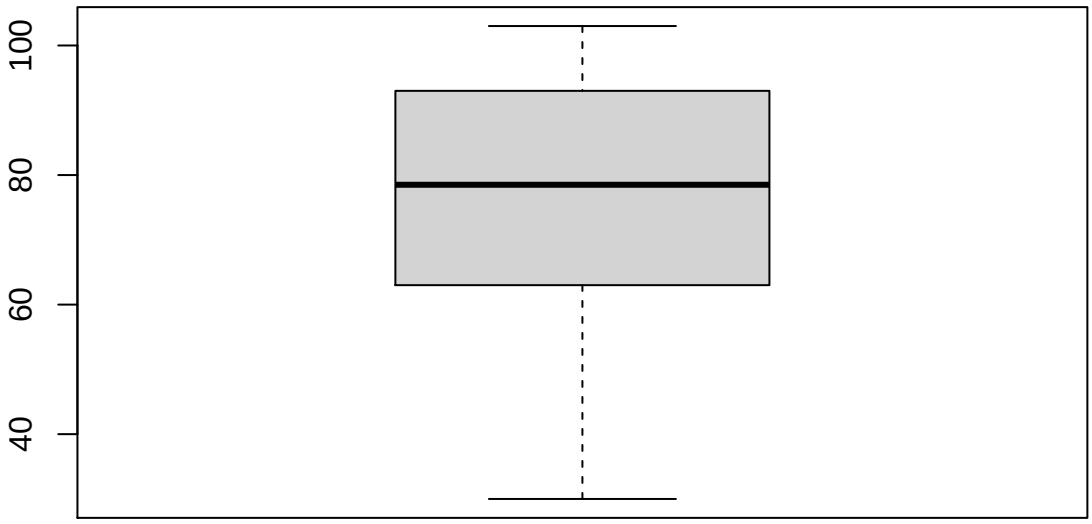
```
summary(df$weeks_alive)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   30.00   63.50   78.50   75.55   92.75  103.00
```

```
sd(df$weeks_alive)
```

```
## [1] 21.42832
```

```
boxplot(df$weeks_alive)
```



Now summary statistics within the different groups. First, I'll collect some basic info using the `tapply()` function:

```
tapply(df$weeks_alive, df$dose, summary)
```

```
## $None
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   70.00   85.00   93.00   91.36  101.00  103.00
##
## $Low
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   49.00   63.00   70.00   69.89   77.00   89.00
##
## $Medium
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   30.00   58.25   79.50   71.50   89.25   97.00
##
## $High
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   34.00   45.00   56.50   65.25   92.75  102.00
```

Here's an alternative way to do this using tidyverse `group_by()` and `summarize_all()` functions:

```
df %>%
  group_by(dose) %>%
  summarize(
    n = n(),
    min = min(weeks_alive),
    avg = round(mean(weeks_alive), 2),
    max = max(weeks_alive),
    sd = round(sd(weeks_alive), 2)
  )
```

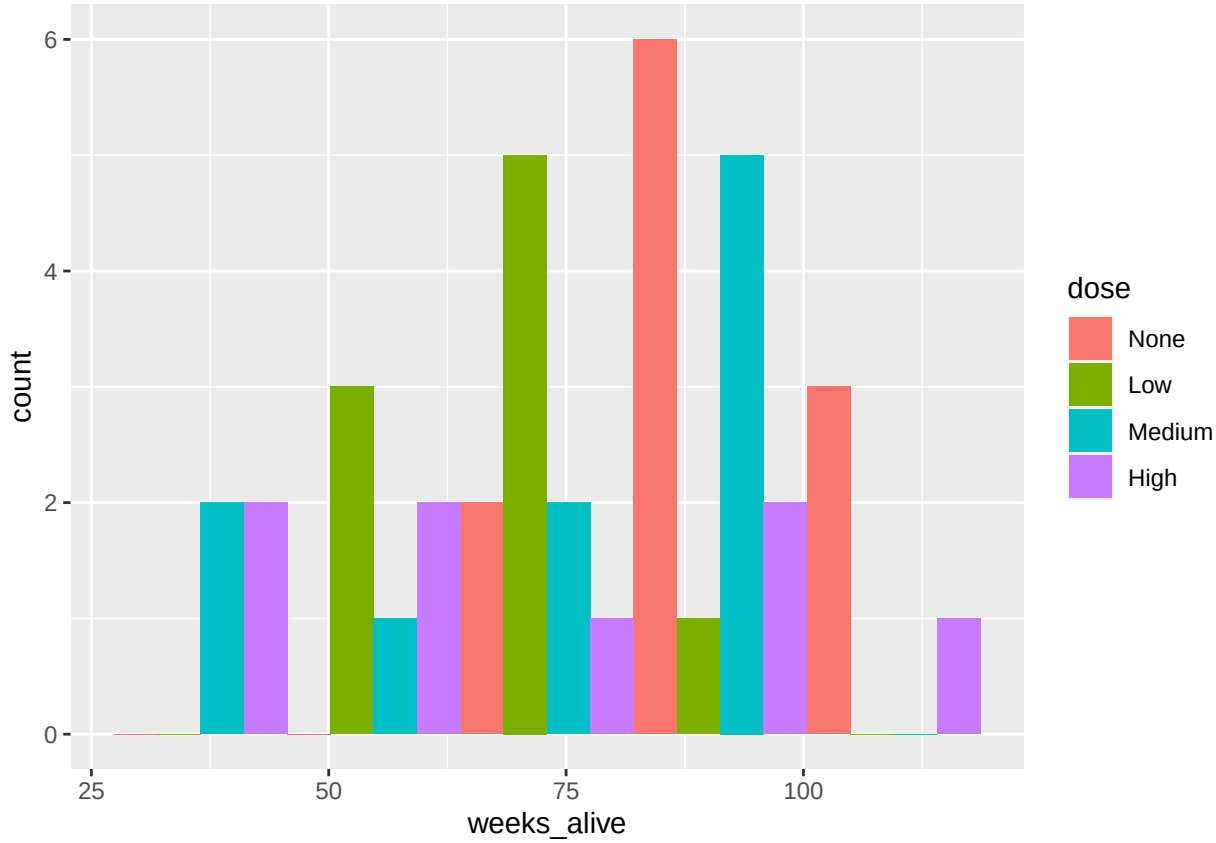
```
## # A tibble: 4 x 6
##   dose       n   min   avg   max    sd
##   <fct>  <int> <dbl> <dbl> <dbl> <dbl>
## 1 None      11    70  91.4   103  11.0
```

```
## 2 Low         9    49  69.9    89  11.9
## 3 Medium     10    30  71.5    97  23.8
## 4 High        8    34  65.2   102  28.1
```

When it comes to visualizations, we definitely want to use ggplot. We have lots of options for what to do.

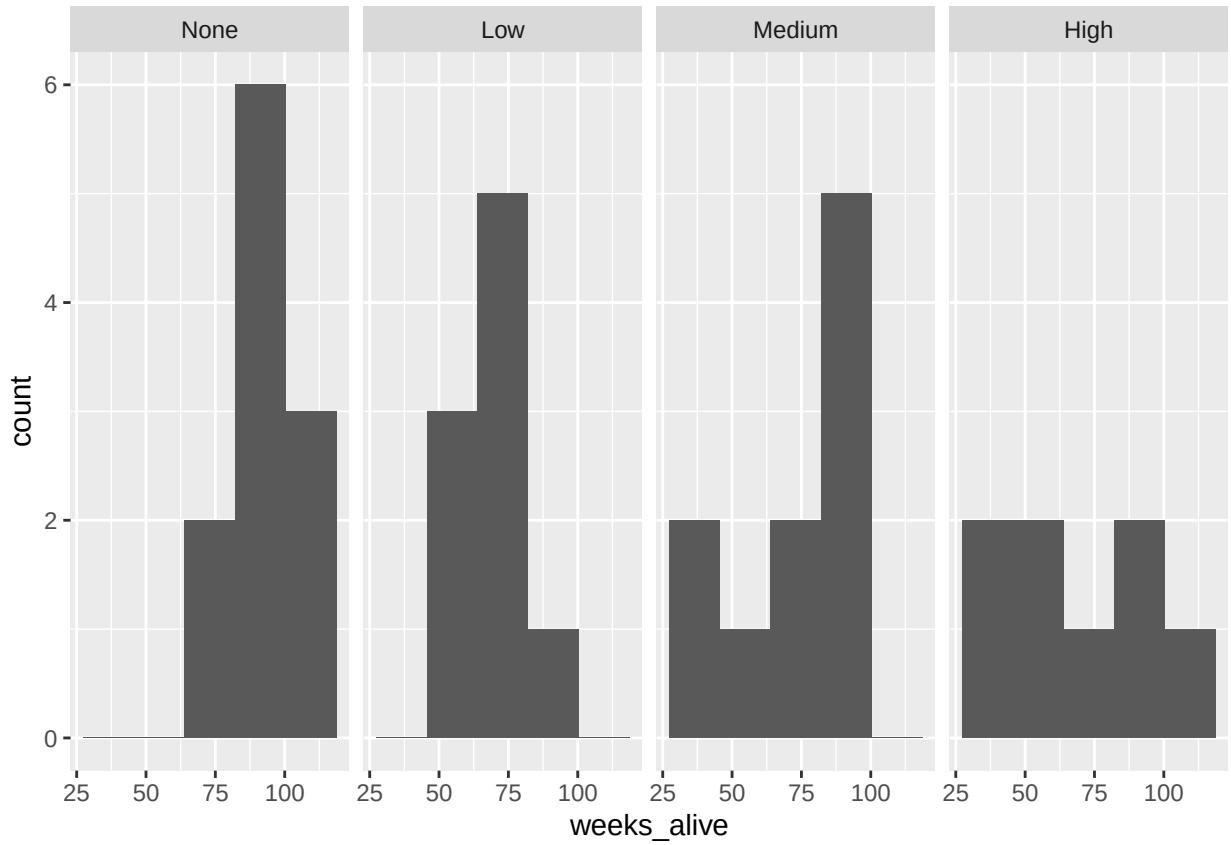
```
library(ggplot2)

# Histograms
h_plot <- ggplot(data = df, aes(x = weeks_alive, fill = dose)) +
  geom_histogram(position = "dodge", bins = 5)
h_plot
```



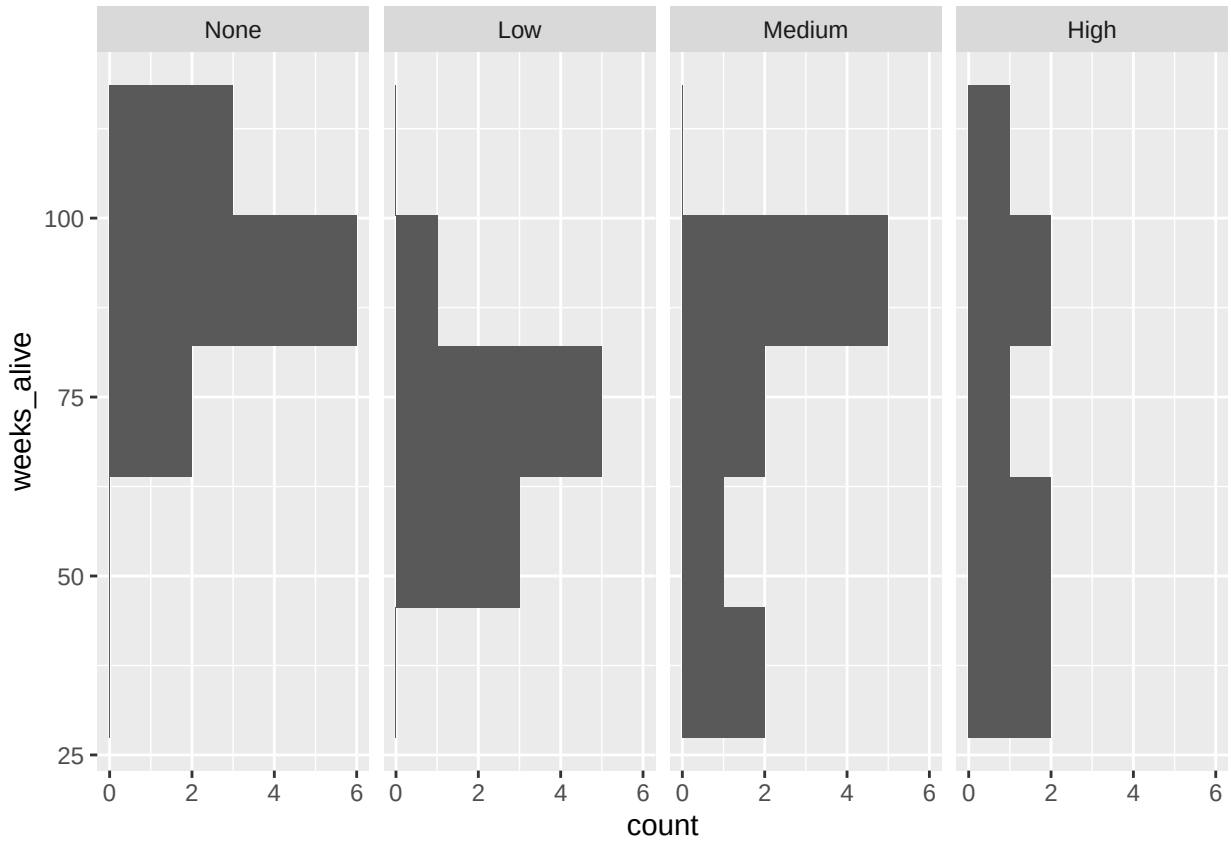
In this case, faceted histograms is probably better

```
h_facet <- ggplot(data = df, aes(x = weeks_alive)) +
  geom_histogram(bins = 5) +
  facet_grid(~dose)
h_facet
```



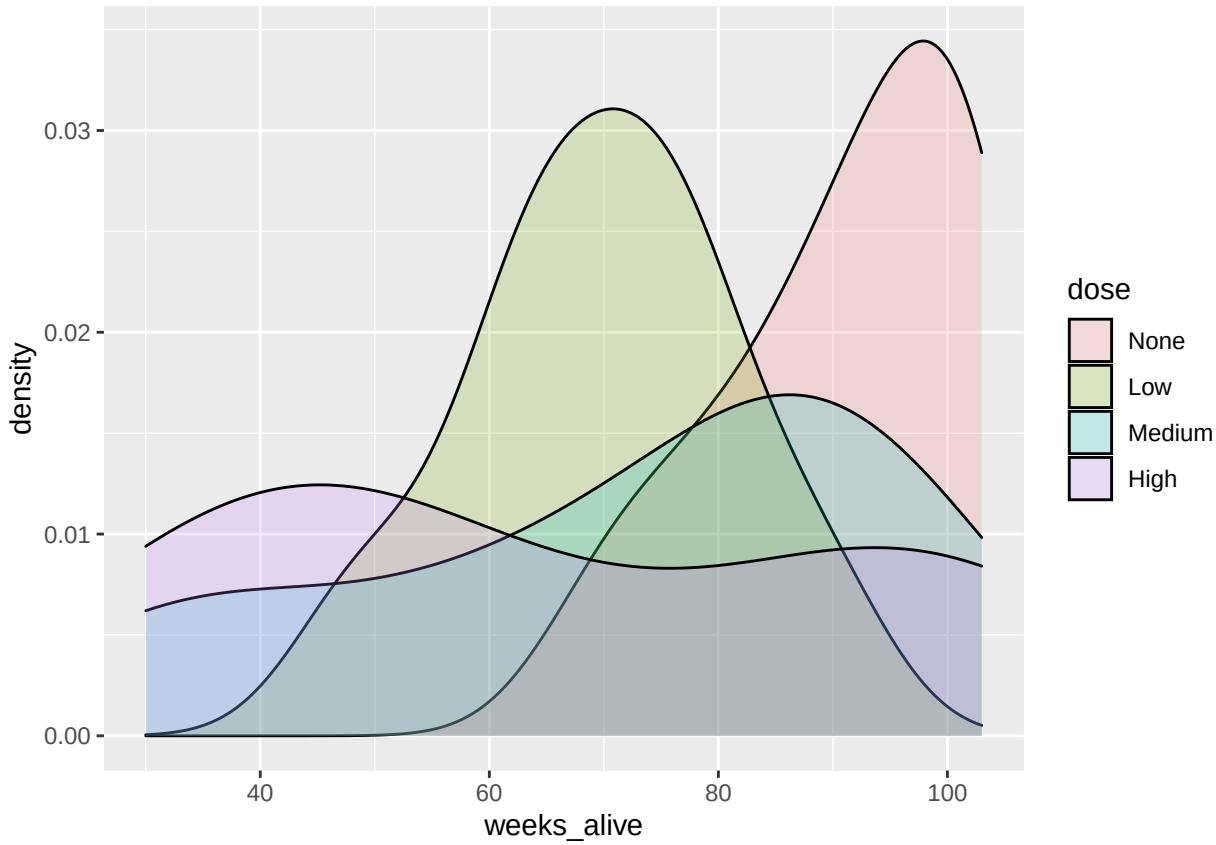
This might be easier to compare if we flip the coordinates (rotate the axes):

```
h_facet + coord_flip()
```



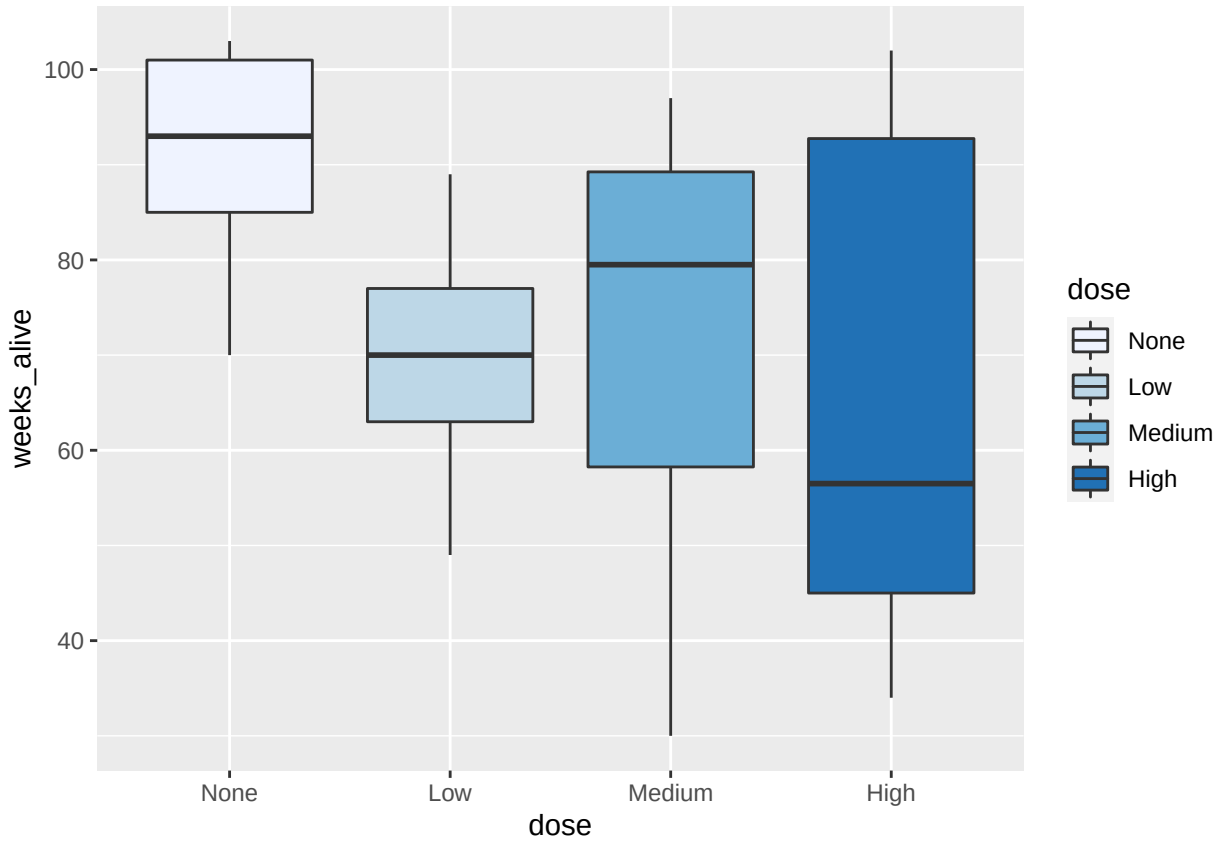
Density plots might be nice? Note that setting a value <1 for "alpha" makes the fill more transparent

```
d_plot <- ggplot(data = df, aes(x = weeks_alive, fill = dose)) +
  geom_density(alpha = .2)
d_plot
```



Aaron finds boxplots quite clarifying and added a call to `scale_fill_brewer()` here to use the "Brewer" color palettes:

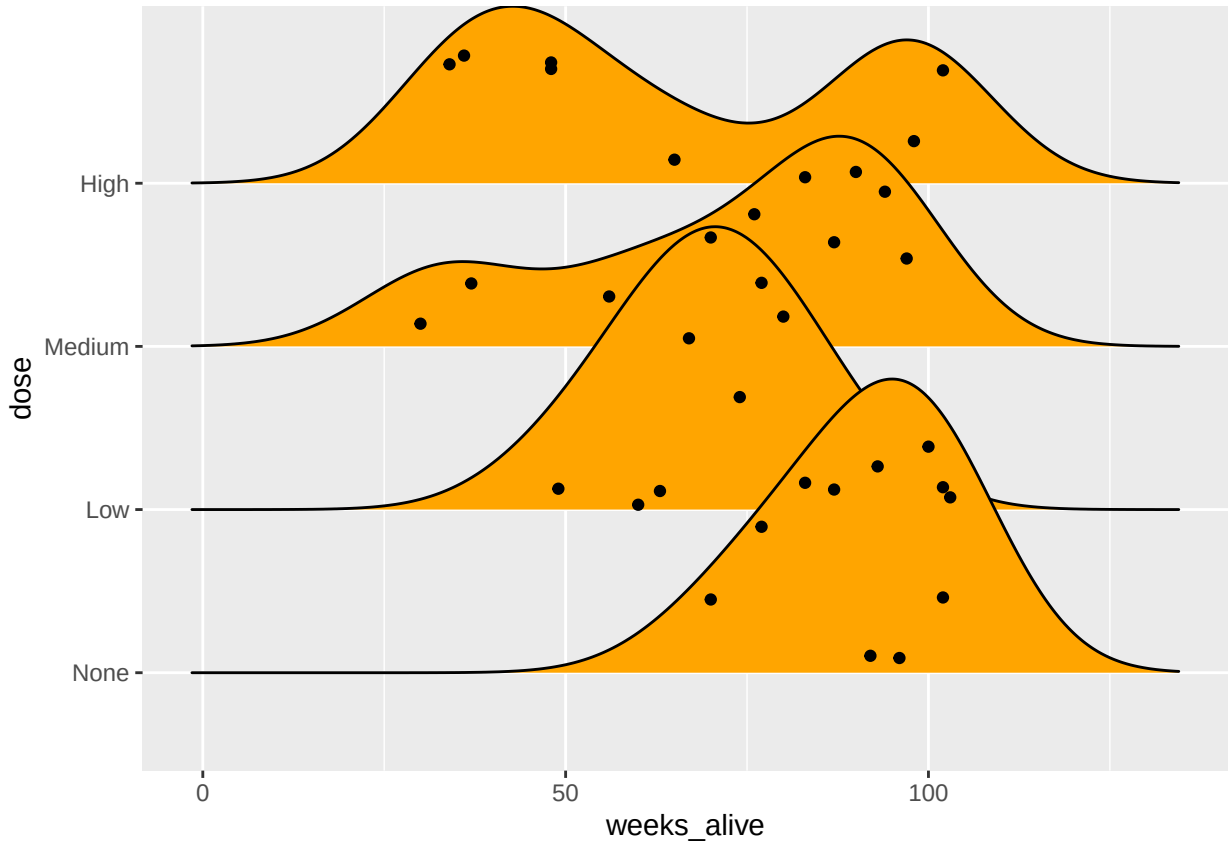
```
box_plot <- ggplot(data = df, aes(y = weeks_alive, x = dose, fill = dose)) +
  geom_boxplot() +
  scale_fill_brewer()
box_plot
```



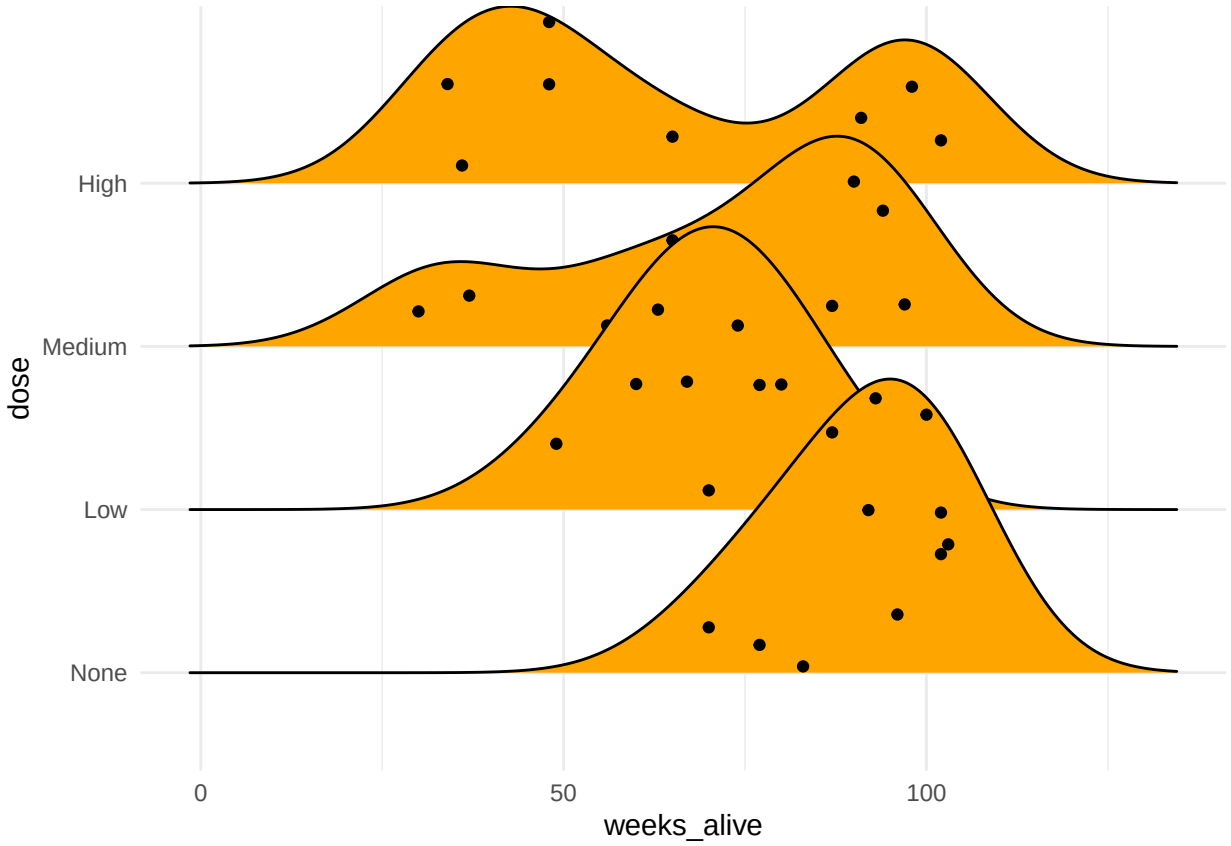
Some people like "ridgeline" plots too:

```
# install.packages('ggridges')
library(ggridges)

ridge_plot <- ggplot(data = df, aes(x = weeks_alive, y = dose)) +
  geom_density_ridges(jittered_points = T, fill = "orange")
ridge_plot
```




```
# add a fancy minimalist theme to make it prettier:
ridge_plot + theme_minimal()
```



SQ1 Discuss the descriptive results Overall, the survival times range from $30 - 100$ weeks, with an average of about 75.5 weeks and a standard deviation of 21.4 weeks. The distribution of outcomes is slightly left-skewed.

Comparisons of survival times across the different dosage levels reveals that the average survival time is highest in the "None" dosage level ($\mu \approx 91$ weeks) and lowest in the "High" dosage level ($\mu \approx 65$ weeks). The spread of the data is also much greater in the "Medium and"High" dosage level groups, suggesting a wider range of outcomes in those conditions. Based on the "ridgeline" plots, it appears that this expanded spread might have something to do with a bi-modal distribution in both conditions.

SQ2 State hypotheses For the ANOVA:

$H_0 : \mu_{none} = \mu_{low} = \mu_{medium} = \mu_{high}$
H_A : Means are not all equal.

For the t-tests of none vs. any:
$H_0 : \mu_{none} = \mu_{any}$
$H_A : \mu_{none} \neq \mu_{any}$

For the t-tests of none vs. high:
$H_0 : \mu_{none} = \mu_{high}$
$H_A : \mu_{none} \neq \mu_{high}$

SQ3 Address assumptions An ANOVA assumes independence, normality, and equal variance. A two sample t-test assumes independence and normality. The assumption of equal variance seems like a stretch (check out those standard deviations within conditions!). Normality is also a bit of a hard sell within groups or overall, but ultimately not so bad. Independence of the samples seems just fine since this was an experiment and we have no reason to anticipate dependence between the different conditions.

Despite the issues with unequal variance and normality, most analysts would march ahead with the analysis despite these violations of assumptions. We can discuss how you might think and talk about this more in class.

PC3 ANOVA analysis

Remember that we have to call `summary()` to see the ANOVA table:

```
summary(aov(weeks_alive ~ dose, data = df))
```

```
##             Df Sum Sq Mean Sq F value Pr(>F)
## dose         3   4052  1350.7    3.55 0.0245 *
## Residuals   34  12937   380.5
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

SQ4 Interpret ANOVA results This provides evidence in support of the alternative hypothesis that the group means are not equal ($p < 0.05$). It seems that the dosage of Red Dye #40 likely has a relationship with how many weeks the mice lived.

PC4 Differences in means

T-test between None and Any, and between None and High.

```
t.test(
  df$weeks_alive[df$dose == "None"], # Samples with no dose
  df$weeks_alive[df$dose != "None"] # Samples with any dose
)
```

```
##
##  Welch Two Sample t-test
##
## data:  df$weeks_alive[df$dose == "None"] and df$weeks_alive[df$dose != "None"]
## t = 4.2065, df = 33.732, p-value = 0.0001806
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  11.49879 33.00626
## sample estimates:
## mean of x mean of y
##  91.36364  69.11111
```

Or, using formula notation

```
t.test(weeks_alive ~ dose == "None", data = df)
```

```
##
##  Welch Two Sample t-test
##
## data:  weeks_alive by dose == "None"
## t = -4.2065, df = 33.732, p-value = 0.0001806
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -33.00626 -11.49879
## sample estimates:
## mean in group FALSE  mean in group TRUE
##            69.11111            91.36364
```

T-test between None and High

```r
t.test(
  df$weeks_alive[df$dose == "None"], # Samples with no dose
  df$weeks_alive[df$dose == "High"] # Samples with high dose
)
```

```
##
##  Welch Two Sample t-test
##
## data:  df$weeks_alive[df$dose == "None"] and df$weeks_alive[df$dose == "High"]
## t = 2.4958, df = 8.5799, p-value = 0.0353
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##   2.266399 49.960874
## sample estimates:
## mean of x mean of y
##  91.36364  65.25000
```

Formula notation for this is a bit tricker. I would create a temporary dataframe...

```r
df.tmp <- subset(df, subset = dose == "None" | dose == "High")
t.test(weeks_alive ~ dose, data = df.tmp)
```

```
##
##  Welch Two Sample t-test
##
## data:  weeks_alive by dose
## t = 2.4958, df = 8.5799, p-value = 0.0353
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##   2.266399 49.960874
## sample estimates:
## mean in group None mean in group High
##           91.36364           65.25000
```

SQ5 Interpret t-test results These t-tests both support the idea that receiving a dose of RD40 reduces lifespan.

SQ6 Multiple comparisons We might not completely trust these p-values, since we are doing multiple comparisons ($\binom{4}{2} = 6$ comparisons, to be precise). One option is to do a Bonferroni correction, where we only cnsider things significant if $\alpha < .05/m$ where m is the number of tests (so our new critical p-value would be 0.0083).

However, as discussed in the Reinhart book, the Bonferroni correction is usually more conservative than it needs to be, and there are other approaches; for example, Benjamini-Hochberg correction which I would calculate calling `pairwise.t.test()` in R like so:

```r
with(
  df,
  pairwise.t.test(weeks_alive, dose, p.adjust = "BH")
)
```

```
##
##  Pairwise comparisons using t tests with pooled SD
##
## data:  weeks_alive and dose
##
```

```
##        None  Low   Medium
## Low    0.052 -     -
## Medium 0.052 0.858 -
## High   0.041 0.753 0.753
##
## P value adjustment method: BH
```

In this particular study, I would suggest reporting adjusted p-values for all of the tests (and I would consider including the "raw" p-values in a footnote or appendix or something).

Empirical paper questions

EQ1 — RQ4 questions

(a) For RQ4, the units of analysis are the 109 respondents. The credibility index (a group of five survey items evaluating how credible the respondents perceived the blogs they read to be) was split at the mean and these high/low groups were the independent (grouping) variable in the ANOVA. The six perceived relationship management factors (survey items) were the dependent variables for the ANOVA.

(b) The analysis tests whether the mean responses on the survey items composing each of the different relationship management factors were equal across the high/low blog credibility groups. The alternative hypotheses are whether there are differences between the groups for any of the perceived relationship management dimensions.

(c) None of the ANOVA tests rejected the null hypothesis of no difference. In other words, there was no evidence that perceptions of relationship management dimensions varied across individuals perceiving blogs as low or high credibiliy.

(d) It is (usually) a bit hard to say much from a null result! See the answer to (c) above.

EQ2 — RQ5 questions

(a) Again, the units are the 109 respondents and the partitioned (low/high) credibility index serves as the independent (grouping) variable. The crisis index is the dependent variable.

(b) The ANOVA tests whether average assessments of perceived crisis in the organization in question were equal by whether participants perceived the blogs to be low/high credibility. The alternative hypotheses are whether there are differences between the groups for perceptions of the organization being in crisis.

(c) The results of the ANOVA reject the null, suggesting support for the alternative hypothesis that participants reporting low credibility blogs reported different (higher) scores on the crisis index.

(d) I find the reported differences compelling, but would like more information in order to determine more specific takeaways. For example, I would like to see descriptive statistics about the various measures to help evaluate whether they meet the assumptions for identifying the ANOVA. Survey indices like this are a bit slippery insofar as they can seem to yield results when the differences are really artifacts of the measurements and how they are coded. I am also a bit concerned that the questions seemed to ask about blog credibility in general rather than the specific credibility of the specific blogs read by the study participants? The presumed relationship between credibility and the assignment to the blogs in question is not confirmed empirically, meaning that the differences in perceptions of organizational crisis might be more related to baseline attitudes than to anything specific about the treatment conditions in the experiment. I would also like to know more about the conditional means and standard errors in order to evaluate whether the pairwise average perceptions of organizational crisis vary across perceived credibility levels.

EQ3 — RQ6 questions

(a) Analogous to RQ5 except that the (six) different dimensions of relationship management separated into high/low categories served as the independent (grouping) variables in the ANOVA. Perceptions of organizational crisis remained the dependent variable.

(b) This set of ANOVAs test whether assessments of perceived organizational crisis were equal or varied depending on the prevalence of specific relationship management strategies.

(c) The results of the ANOVAs are mixed, rejecting the null hypothesis of no difference for two of the relaionship management dimensions (task sharing and responsiveness/customer service), but failing to do so for the other four dimensions. For the two dimensions in question, higher prevalence of each strategy appeared to reduce the perception of crisis.

(d) Again, the differences reported are compelling insofar as they indicate larger varation across the groups in terms of perceived crisis than would be expected in a world where no difference existed. That said, in addition to all the issues mentioned above for EQ5 part (d), this set of tests also raisesw issues of multiple comparisons. Not only is it useful to consider multiple comparisons in the context of a single ANOVA, but when running many ANOVAs across many grouping variables. If the authors really wanted to test whether the specific PR strategies mentioned here altered perceptions of organizational crisis across different types of blogs, they should/could have incorporated those variations into their experimental design much more directly. As such, the results remain suggestive that some of the observed relationships may exist, but can provide only weak support for even those that reject the null hypotheses in statistical tests.