Project 1 Exploratory Analysis

Jack Cunningham

Loading packages and data

```
library(tidyverse)
```

```
-- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
v dplyr     1.1.4     v readr     2.1.5
v forcats   1.0.0     v stringr   1.5.1
v ggplot2   3.5.2     v tibble    3.2.1
v lubridate 1.9.4     v tidyr     1.3.1
v purrr     1.0.4     
-- Conflicts ------------------------------------------ tidyverse_conflicts() --
x dplyr::filter() masks stats::filter()
x dplyr::lag()    masks stats::lag()
i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to becon
```

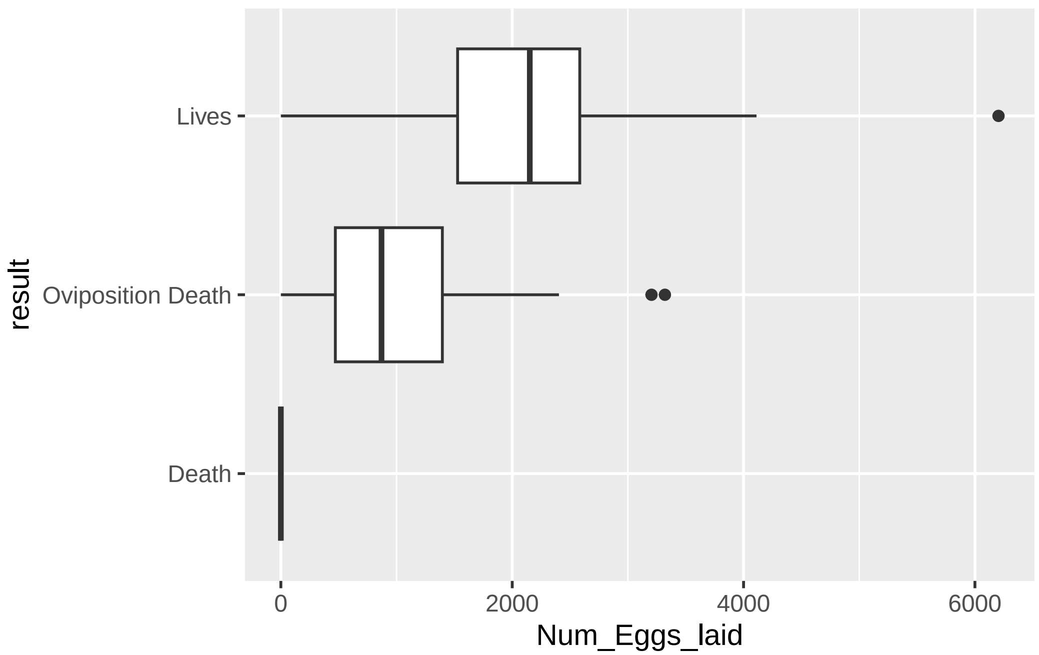
```
load("data_processed.RData")
```

Response variables

Starting with numeric variables.

Starting first with eggs laid:

```
data_processed |>
  ggplot(aes(Num_Eggs_laid,result)) +
  geom_boxplot()
```



We can see that the number of eggs laid is a steady zero for when the tick dies. If the tick dies during oviposition the number of eggs laid is generally lower than when the tick lives. The variation in the number of eggs is also greater when the tick lives, compared to oviposition death. There exists one outlier value from the number of eggs being laid exceeds 6000. Lets look at that row:

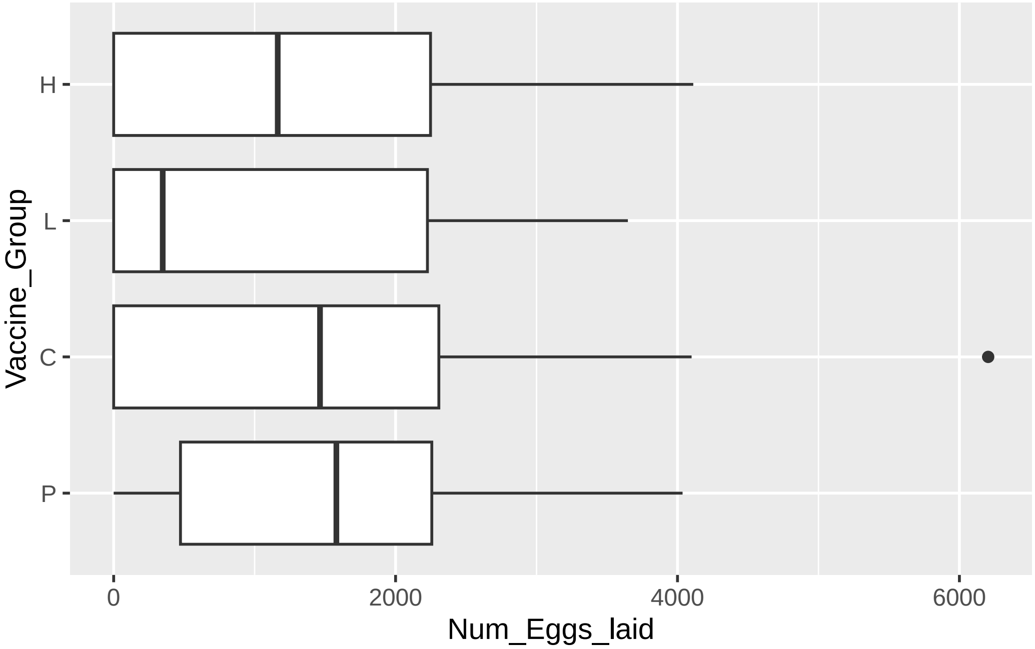
```
data_processed |>
  filter(Num_Eggs_laid > 6000)
```

```
# A tibble: 1 x 14
  Tick_ID Deer_ID Vaccine_Group Infestation Mortality `Weight_Replete_Female(g)`
  <chr>   <fct>   <fct>         <fct>       <fct>                          <dbl>
1 793_67  793     C             FIRST       0                              0.262
# i 8 more variables: `Weight_ Egg_Mass(g)` <dbl>,
#   `Bloodmeal_Conversion(%)` <dbl>, Num_Eggs_laid <dbl>, Num_Larvae <dbl>,
#   Comments <chr>, ovi_position <lgl>, result <fct>, percent_hatched <dbl>
```

There does not appear to be anything suspicious about this observation, and it may indicate some right skew in the number of eggs laid variable. Generally, this is a symmetric distribution however.

Let's look at the eggs laid across the vaccination groups:

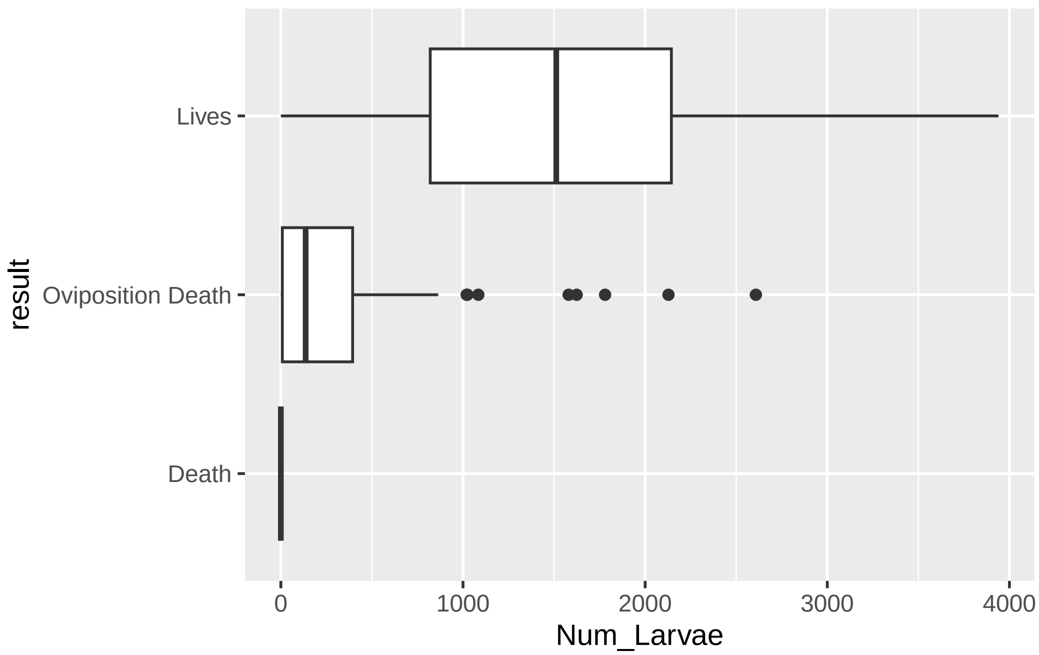
```
data_processed |>
  ggplot(aes(Num_Eggs_laid,Vaccine_Group)) +
  geom_boxplot()
```



Interestingly enough it appears that when the vaccine dose is low the lower number of eggs are laid. The control and the higher vaccine dose seem quite similar when it comes to this metric.

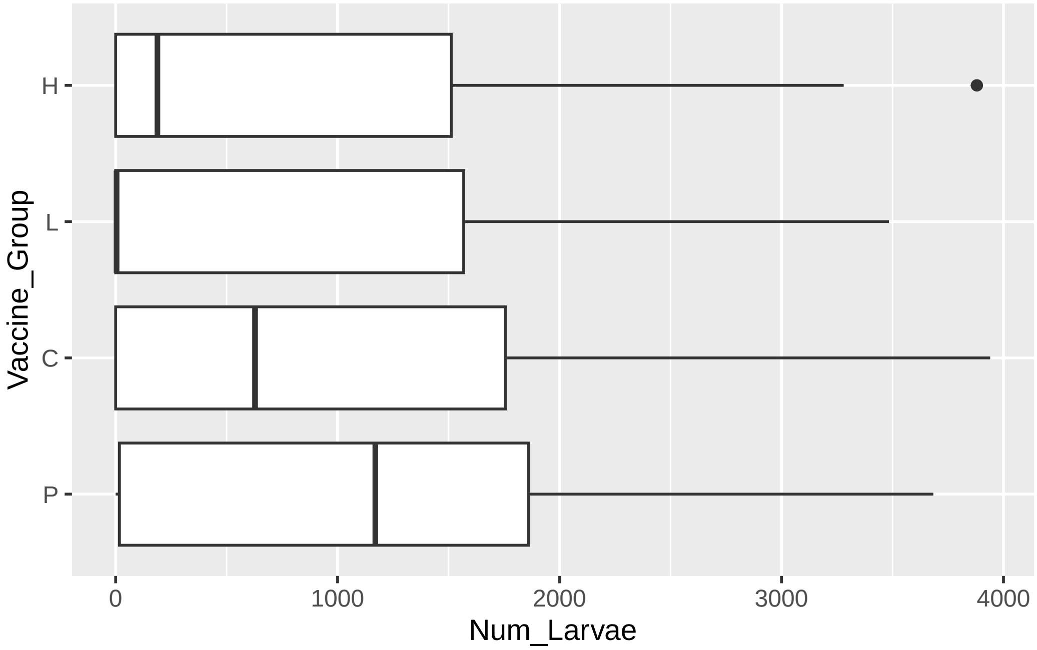
This motivates me to take a look at another response variable we have, number of larvae. Lets see how this varies across vaccine groups:

```
data_processed |>
  ggplot(aes(Num_Larvae,result)) +
  geom_boxplot()
```



We can see that when the result is Oviposition death there is a lot less larvae than when the result is lives. The distribution of the number of Larvae when oviposition is death is very right skewed however, occasionally there are many Larvae despite the Oviposition Death.

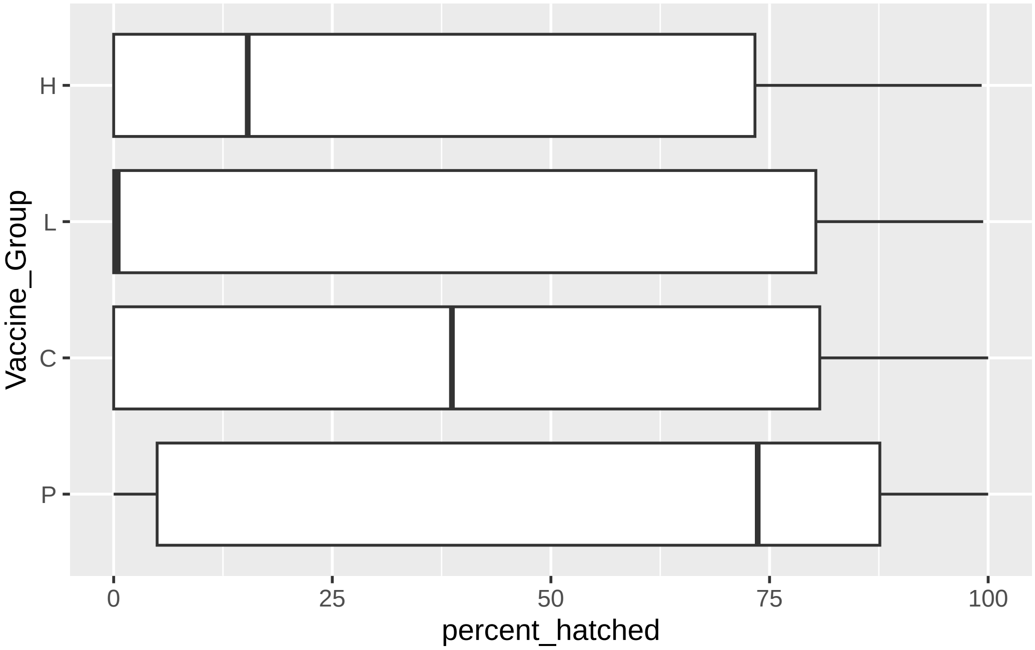
```
data_processed |>
  ggplot(aes(Num_Larvae,Vaccine_Group)) +
  geom_boxplot()
```



We can see that the number of larvae is much lower in the vaccination groups when compared to the two non vaccination groups.

Let's look at % Hatched:

```
data_processed |>
  ggplot(aes(percent_hatched,Vaccine_Group)) +
  geom_boxplot()
```

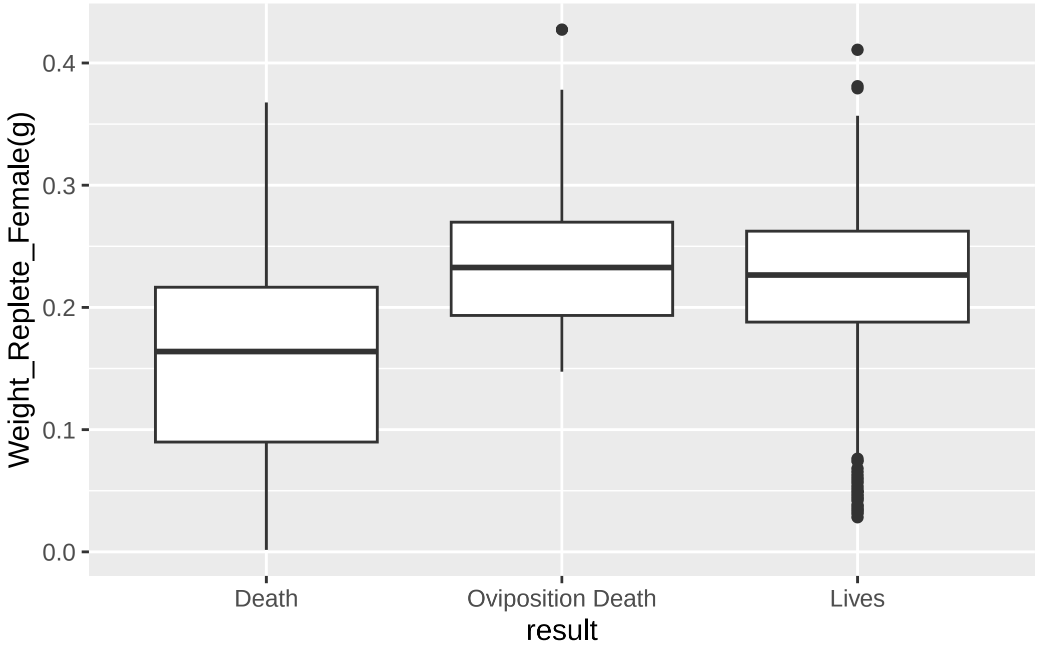


We can see that the vaccination groups L and H both have percent hatched smaller in general, but each distribution is disperse across the full range of reasonable percentages.

Covariates

Lets see if Weight_Replete_Female has a strong relationship with the result:

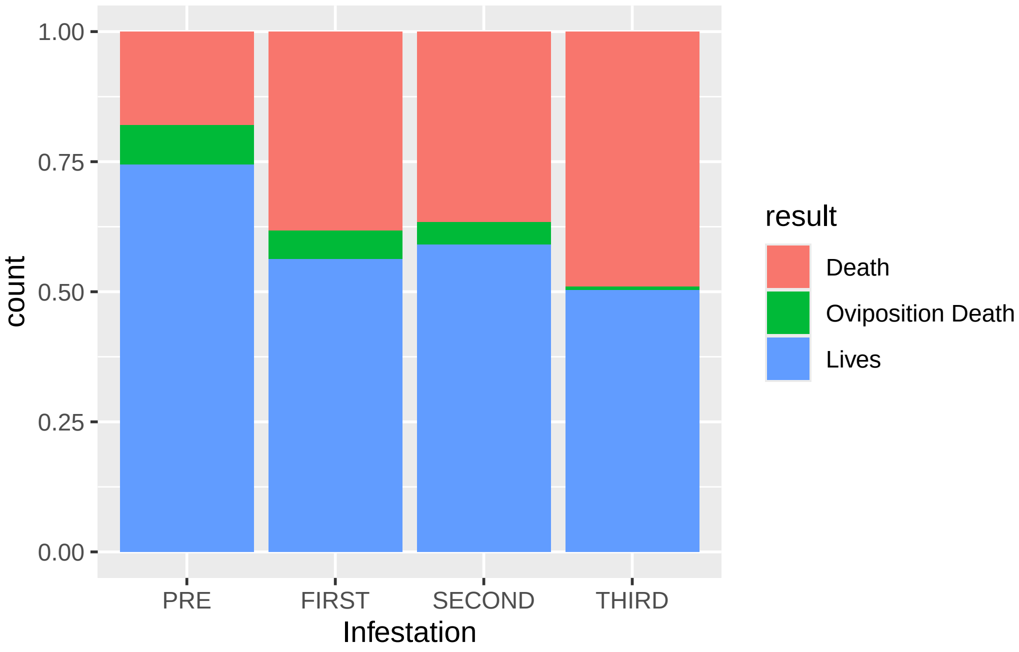
```
data_processed |>
  ggplot(aes(result,`Weight_Replete_Female(g)`)) +
  geom_boxplot()
```



The difference doesn't seem that extreme.

Lets see if Infestation is strongly related to the result:

```
data_processed |>
  ggplot(aes(fill = result, x = Infestation)) +
  geom_bar(position = "fill")
```

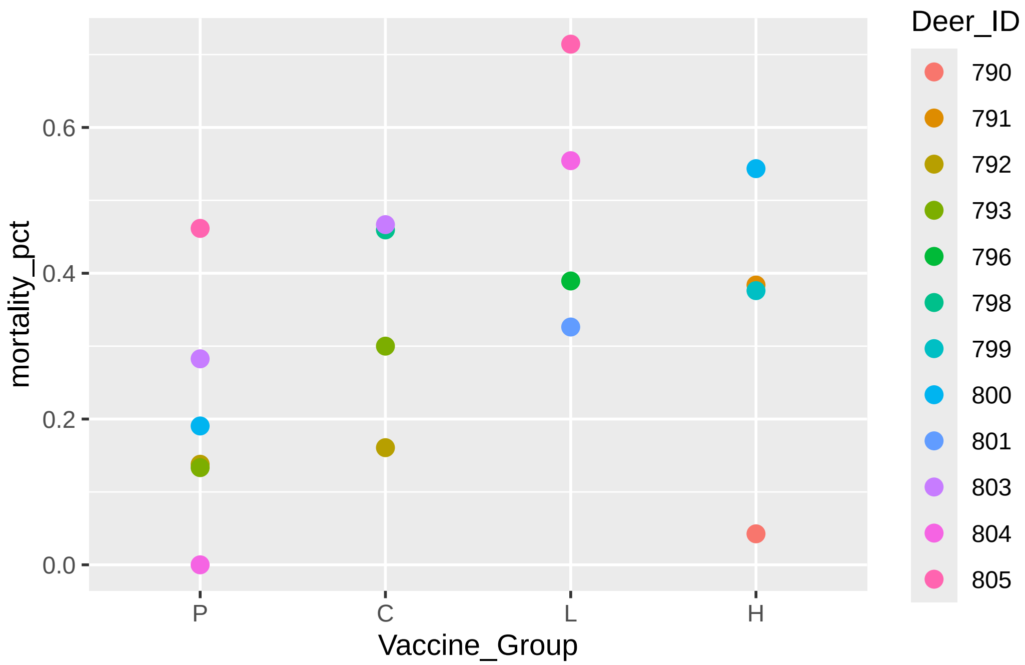


Considering that the pre infestation is confounded with the fact that there were no vaccines administered, the time of infestation does not seem to be an important covariate when it comes to the result.

Lets take a look at the variation within the groups of vaccination, this will give us an idea of what impact the deer is having.

```
data_processed |>
  group_by(Deer_ID,Vaccine_Group) |>
  summarise(mortality_pct = mean(Mortality == 1)) |>
  ggplot(aes(x = Vaccine_Group, y = mortality_pct, col = Deer_ID)) +
  geom_point(shape = 16
             , size = 3)
```

```
`summarise()` has grouped output by 'Deer_ID'. You can override using the
`.groups` argument.
```



We can see that there appears to be pretty large variation between deer in each vaccination group, this may make it more difficult to prove that vaccines are making the impact versus variation between deer.