Numerical Data Applications Solutions

Professor Bradley Warner

27 August, 2020

Exercises

Create this Rmd file for the work including headers, file creation data, and explanation of your work. Make sure your plots have a title and the axes are labeled. We are asking you to do more in this application to get ready for your Oral Board.

1. **Mammals exploratory**

Data were collected on 39 species of mammals distributed over 13 orders. The data is in the `openintro` package as `mammals`

a. Using help, report the units for the variable `BrainWt`.

```
?mammals
```

b. Using `inspect` how many variables are numeric?

```
inspect(mammals)
```

```
## Warning: ‘data_frame()‘ is deprecated as of tibble 1.1.0.
## Please use ‘tibble()‘ instead.
## This warning is displayed once every 8 hours.
## Call ‘lifecycle::last_warnings()‘ to see where this warning was generated.
```

```
##
## categorical variables:
##      name  class levels  n missing
## 1 species factor     62 62       0
##                                    distribution
## 1 Africanelephant (1.6%) ...
##
## quantitative variables:
##               name   class   min     Q1  median       Q3    max       mean
## ...1       body_wt numeric  0.005  0.600  3.3425  48.2025 6654.0 198.789984
## ...2      brain_wt numeric  0.140  4.250 17.2500 166.0000 5712.0 283.134194
## ...3  non_dreaming numeric  2.100  6.250  8.3500  11.0000   17.9   8.672917
## ...4      dreaming numeric  0.000  0.900  1.8000   2.5500    6.6   1.972000
## ...5   total_sleep numeric  2.600  8.050 10.4500  13.2000   19.9  10.532759
## ...6     life_span numeric  2.000  6.625 15.1000  27.7500  100.0  19.877586
## ...7     gestation numeric 12.000 35.750 79.0000 207.5000  645.0 142.353448
## ...8     predation integer  1.000  2.000  3.0000   4.0000    5.0   2.870968
## ...9      exposure integer  1.000  1.000  2.0000   4.0000    5.0   2.419355
## ...10       danger integer  1.000  1.000  2.0000   4.0000    5.0   2.612903
##                 sd  n missing
## ...1  899.158011 62       0
## ...2  930.278942 62       0
## ...3    3.666452 48      14
## ...4    1.442651 50      12
## ...5    4.606760 58       4
## ...6   18.206255 58       4
## ...7  146.805039 58       4
## ...8    1.476414 62       0
## ...9    1.604792 62       0
## ...10   1.441252 62       0
```

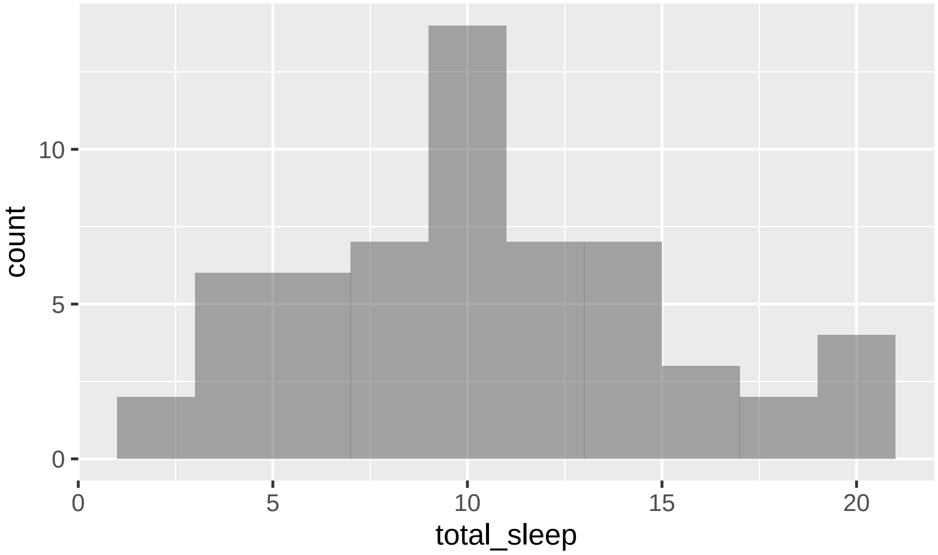
c. What type of variable is `danger`?

Categorical

d. Create a histogram of `total_sleep` and describe the distribution.

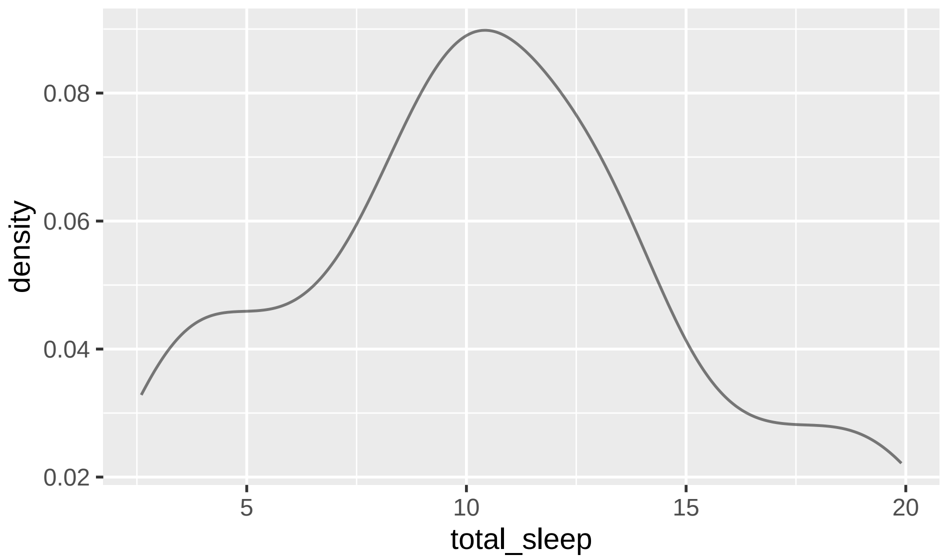
```
gf_histogram(~total_sleep,data=mammals,binwidth = 2)
```

```
## Warning: Removed 4 rows containing non-finite values (stat_bin).
```



```
gf_dens(~total_sleep,data=mammals)
```

```
## Warning: Removed 4 rows containing non-finite values (stat_density).
```

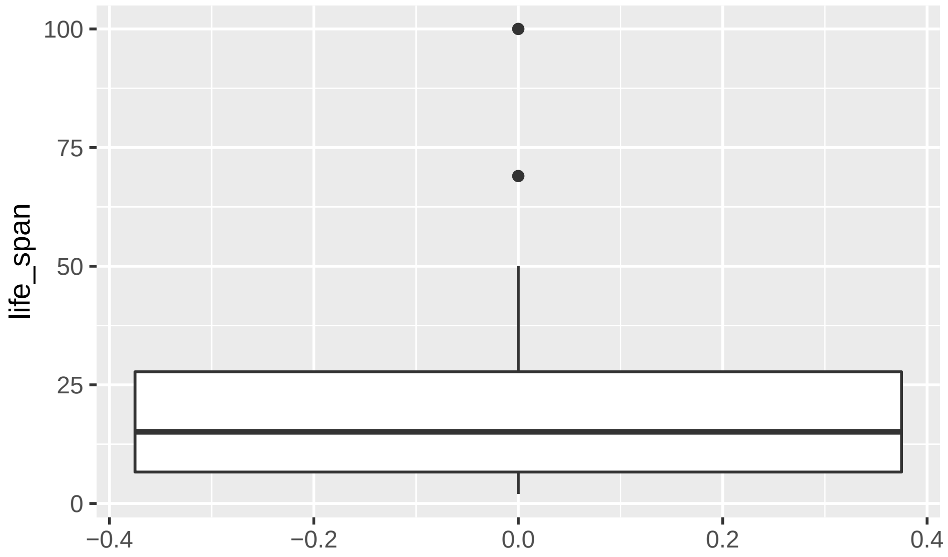


The distribution is unimodal and skewed to the right. It appears it is centered around the value of 11.

e. Create a boxplot of `life_span` and describe the distribution.

```
gf_boxplot(~life_span,data=mammals)
```

```
## Warning: Removed 4 rows containing non-finite values (stat_boxplot).
```



f. Report the mean and median life span of a mammal.

```
mean(~life_span,data=mammals,na.rm=TRUE)
```

```
## [1] 19.87759
```

```
median(~life_span,data=mammals,na.rm=TRUE)
```

```
## [1] 15.1
```

g. Calculate the summary statistics for `LifeSpan` broken down by `Danger`.

```
favstats(life_span~danger,data=mammals)
```

```
##   danger  min     Q1 median     Q3   max     mean       sd  n missing
## 1      1  3.0  7.700  17.60 32.500 100.0 24.20556 23.53829 18       1
## 2      2  2.3  4.500  10.40 13.000  50.0 12.92308 13.15948 13       1
## 3      3  2.0  4.175   5.35  7.875  38.6  9.43750 11.99559  8       2
## 4      4  2.6  9.775  22.10 27.000  69.0 23.11000 18.75482 10       0
## 5      5 17.0 20.000  23.60 30.000  46.0 26.95556 10.18910  9       0
```

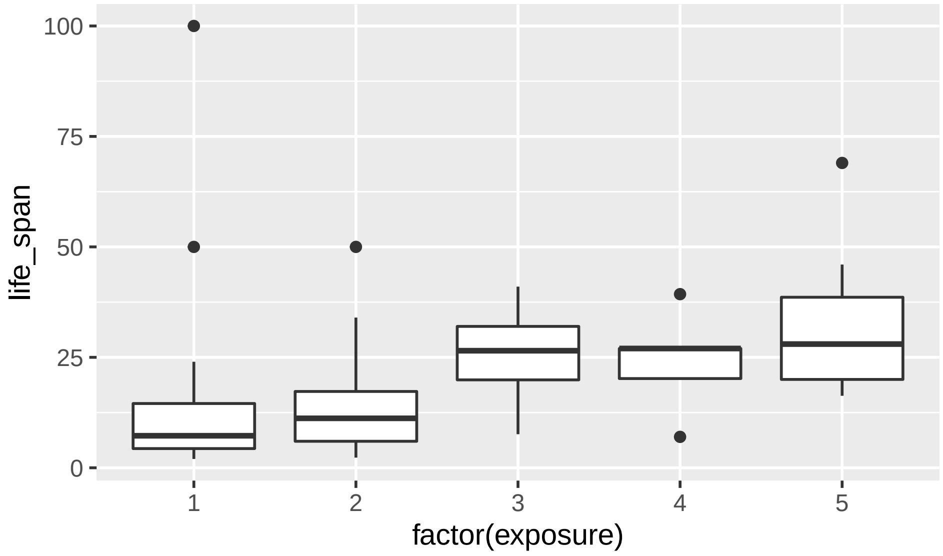
2. **Mammals life spans**

Continue using the `mammals` data set.

a. Create side-by-side boxplots for `life_span` broken down by `exposure`. Note: you will have to change `exposure` to a `factor()`. Report on any findings.

```
mammals %>%
gf_boxplot(life_span~factor(exposure))
```

```
## Warning: Removed 4 rows containing non-finite values (stat_boxplot).
```



b. What happened to the median and third quartile in exposure group 4?

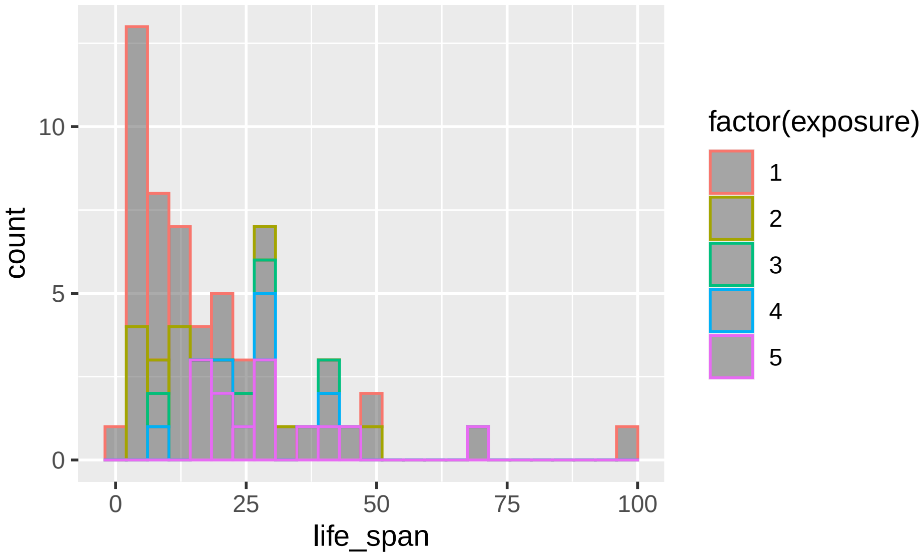
```
favstats(life_span~factor(exposure),data=mammals)
```

```
##   factor(exposure)  min    Q1 median     Q3   max     mean       sd  n missing
## 1                1  2.0  4.35   7.25 14.550 100.0 14.55000 20.98594 24       3
## 2                2  2.3  6.00  11.20 17.275  50.0 15.39167 14.55819 12       1
## 3                3  7.6 19.90  26.50 32.000  41.0 25.40000 13.84582  4       0
## 4                4  7.0 20.20  27.00 27.000  39.3 24.10000 11.78431  5       0
## 5                5 16.3 20.00  28.00 38.600  69.0 30.53077 14.98084 13       0
```

c. Create faceted histograms. What are the shortcomings of this plot?

```
gf_histogram(~life_span,color=~factor(exposure),data=mammals)
```

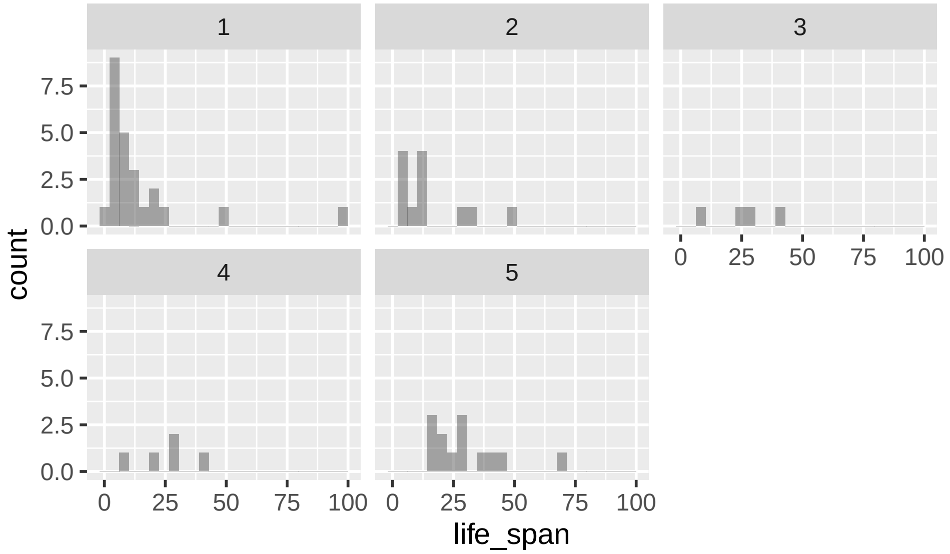
```
## Warning: Removed 4 rows containing non-finite values (stat_bin).
```



This is awful.

```
gf_histogram(~life_span|factor(exposure),data=mammals)
```

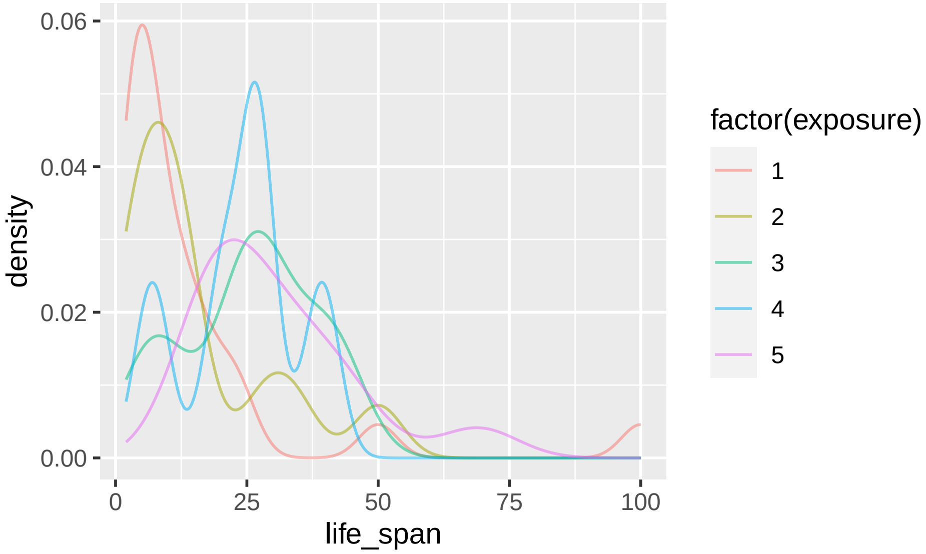
```
## Warning: Removed 4 rows containing non-finite values (stat_bin).
```



Not enough data for each histogram; some of the histograms provide little to no information. Let's do denisty plots.

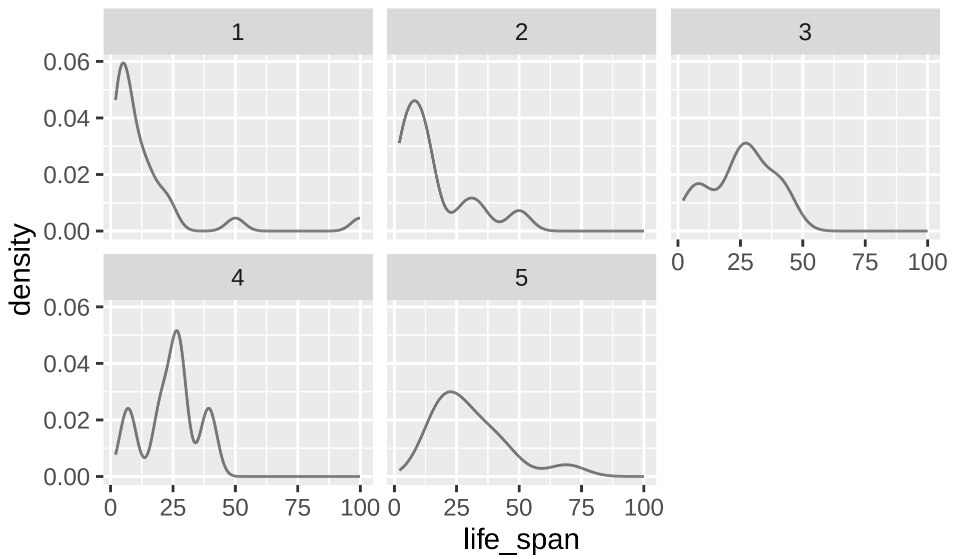
```
gf_dens(~life_span,color=~factor(exposure),data=mammals)
```

```
## Warning: Removed 4 rows containing non-finite values (stat_density).
```



```
gf_dens(~life_span|factor(exposure),data=mammals)
```

```
## Warning: Removed 4 rows containing non-finite values (stat_density).
```



Which do you think is the best graph?

d. Create a new variable `exposed` that is a factor with level `Low` if exposure is `1` or `2` and `High` otherwise.

```
mammals <- mammals %>%
  mutate(exposed=factor(ifelse((exposure==1)|(exposure==2),"Low","High")))
```

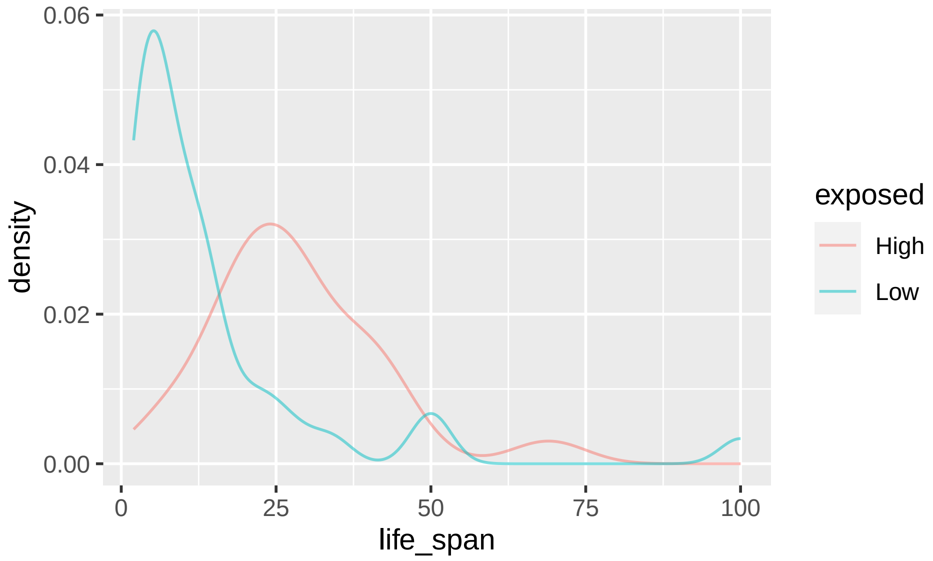
```
inspect(mammals)
```

```
##
## categorical variables:
##      name  class levels  n missing
## 1 species factor     62 62       0
## 2 exposed factor      2 62       0
##                                  distribution
## 1 Africanelephant (1.6%) ...
## 2 Low (64.5%), High (35.5%)
##
## quantitative variables:
##                name   class    min     Q1   median       Q3    max       mean
## ...1       body_wt numeric  0.005  0.600   3.3425  48.2025 6654.0 198.789984
## ...2      brain_wt numeric  0.140  4.250  17.2500 166.0000 5712.0 283.134194
## ...3  non_dreaming numeric  2.100  6.250   8.3500  11.0000   17.9   8.672917
## ...4      dreaming numeric  0.000  0.900   1.8000   2.5500    6.6   1.972000
## ...5   total_sleep numeric  2.600  8.050  10.4500  13.2000   19.9  10.532759
## ...6     life_span numeric  2.000  6.625  15.1000  27.7500  100.0  19.877586
## ...7     gestation numeric 12.000 35.750  79.0000 207.5000  645.0 142.353448
## ...8     predation integer  1.000  2.000   3.0000   4.0000    5.0   2.870968
## ...9      exposure integer  1.000  1.000   2.0000   4.0000    5.0   2.419355
## ...10       danger integer  1.000  1.000   2.0000   4.0000    5.0   2.612903
##               sd  n missing
## ...1  899.158011 62       0
## ...2  930.278942 62       0
## ...3    3.666452 48      14
## ...4     1.442651 50       12
## ...5     4.606760 58        4
## ...6    18.206255 58        4
## ...7   146.805039 58        4
## ...8     1.476414 62        0
## ...9     1.604792 62        0
## ...10    1.441252 62        0
```

e. Repeat part c with the new variable.

```
gf_dens(~life_span,color=~exposed,data=mammals)
```

```
## Warning: Removed 4 rows containing non-finite values (stat_density).
```

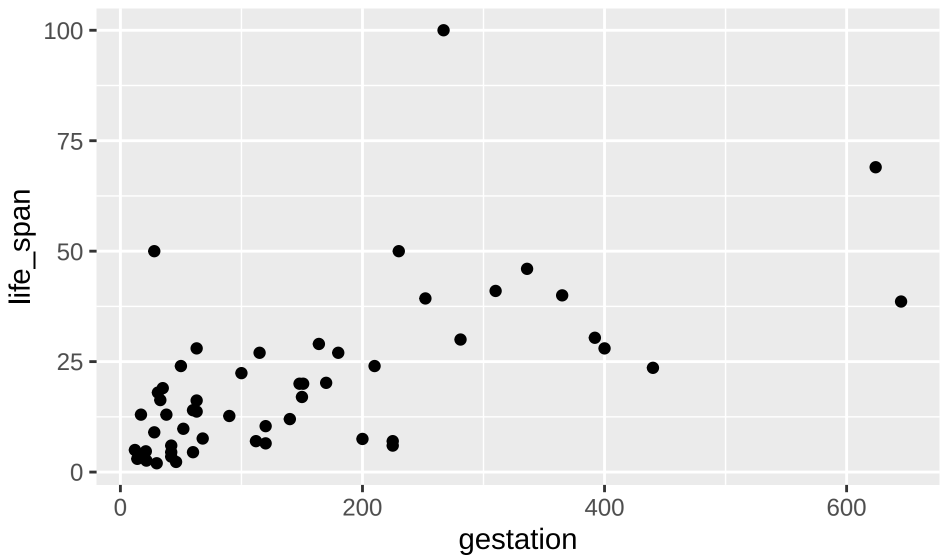


3. **Mammals life spans continued**

a. Create a scatterplot of life span versus length of gestation.

```
mammals %>%
gf_point(life_span~gestation)
```

```
## Warning: Removed 7 rows containing missing values (geom_point).
```



b. What type of an association is apparent between life span and length of gestation?

It is a weak positive association.

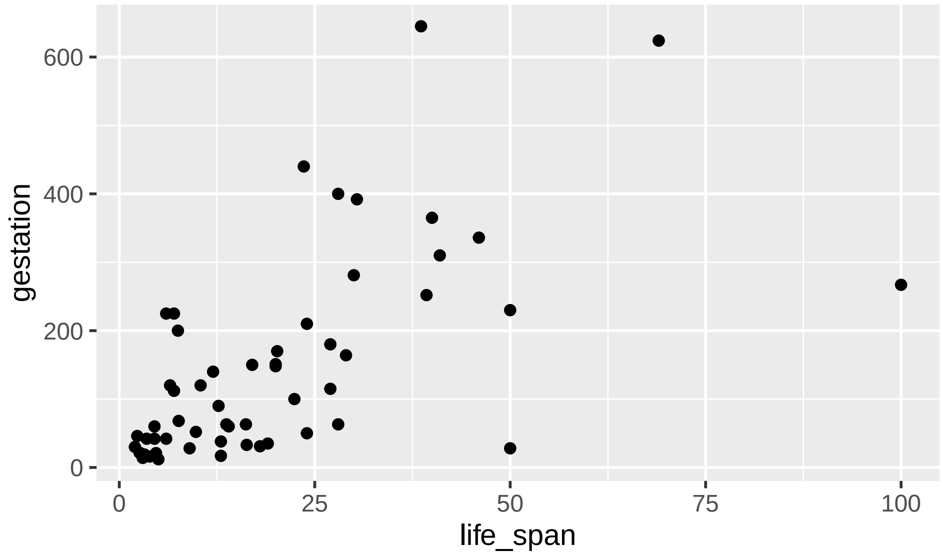
c. What type of an association would you expect to see if the axes of the plot were reversed, i.e. if we plotted length of gestation versus life span?

The same as this is observational data there is no reason to beliee is a causal relationship just by looking at the data. Switching the axis will preserve the association.

d. Create the new scatterplot suggested in c.

```
mammals %>%
gf_point(gestation~life_span)
```

```
## Warning: Removed 7 rows containing missing values (geom_point).
```



e. Are life span and length of gestation independent? Explain your reasoning.

No there is an association and it appears to be linear. If the plot looked like a “shotgun” blast, we would consider the variables to be independent. However, remember there may be confounding variables that could impact the association between these variables.