Cap Dap Final

May Buckingham

11/27/2022

Overview

Here are the three hypotheses that I would like to explore for my data analysis capstone project.

1. I would like to look at the relationship between nest box success rate for kestrels and years that the box is up. I would hypothesize that boxes with more years up would have a higher nest box success rate for kestrels because there would be a longer period of time available for kestrels to nest in the box. Since there would be more time for kestrels to nest, they would have a better chance at successfully fledging.
2. I'd like to explore how many years a nest box needs to be up before it would start to yield kestrel chicks. I believe that the number of years a nest box would need to be up before yielding chicks would be possibly around 3-4 years because that would give the kestrels enough time to find and recognize the nest box as a shelter option. Newer nest boxes might have less kestrel chicks because it may be harder for kestrels to find newer boxes.
3. I would like to look at the relationship between number of chicks banded and success rate for the nest boxes. My question is does higher number of chicks banded result in a higher success rate for the boxes? I would hypothesize that nest boxes with a higher number of chicks banded would result in a higher success rate for the boxes.

Citations:

Smallwood, J. A., Causey, M. F., Mossop, D. H., Klucsarits, J. R., Robertson, B., Robertson, S., Mason, J., Maurer, M. J., Melvin, R. J., Dawson, R. D., Bortolotti, G. R., Parrish, J. W., Breen, T. F., & Boyd, K. (2009). Why are American kestrel (Falco sparverius) populations declining in North America? evidence from nest-box programs. Journal of Raptor Research, 43(4), 274–282. https://doi.org/10.3356/jrr-08-83.1

J.A. Fargallo , G. Blanco , J. Potti & J. Viñuela (2001) Nestbox provisioning ina rural population of Eurasian Kestrels: breeding performance, nest predation and parasitism,Bird Study, 48:2, 236-244, DOI: 10.1080/00063650109461223

Analyzing American Kestrel Nest Box Success

This data set contains data collected by Kealey Viglielmo for a study on American Kestrel nesting box success by Susan Willson, James R. Chandler, Carol Cady, and Mark Manske. The specific data set that I am using contains data of 154 nesting boxes from the years 2002 to 2013. The data contains the number of years the box is up, percent nest success, number of chicks banded, and box inhabitants.

First step is importing packages and clearing R's brain.

Then importing the data set I will be using.

Lets do a summary of all of the data to get an overview of the data and to see what we might have to fix.

Tidying up the data

Lets pivot the years from separate columns into one long column using pivot_longer.

```
kestrels_long <- pivot_longer(kestrels, cols = X2002_inhabitants:X2013_inhabitants, names_to = "year", v
```

lets get rid of X and _inhabitants with string R to tidy up the data:

```
kestrels_long1 <-str_replace(kestrels_long$year, "X", "")

kestrels_long$newYear <- kestrels_long1
view(kestrels_long)
```

```
kestrels_long <- select(kestrels_long, - year)
```

```
kestrels_long1 <- str_replace(kestrels_long$newYear, "_inhabitants", "")
kestrels_long$Year <- kestrels_long1
view(kestrels_long)
```

```
kestrels_long <- select(kestrels_long, -newYear)
view(kestrels_long)
```

Tidy up kestrel variable in inhabitants column.

```
Kestrels <-which(kestrels_long$inhabitants == "KESTRELS" | kestrels_long$inhabitants == "Kestrels" | kes
kestrels_long$inhabitants[Kestrels] <- "KESTRELS"
```

Then I will create a new data set that contains only the data that has Kestrels in the inhabitants variable.

```
kestrel_filter <- filter(kestrels_long, inhabitants == "KESTRELS")
```

Then I will create a new data set that groups the data by the nest box number.

```
Kestrel_summary <- kestrel_filter %>%
  group_by(BOX) %>%
  summarise(
    Years.Up = max(Years.Up),
    Success = max(X.Successful),
    Chicks = max(with_chicks))
```

Then change BOX to a factor because the box number is the identity of the box, not a continuous variable.

```
Kestrel_summary$BOX <- as.factor(Kestrel_summary$BOX)
```

Hypothesis #1

Exploring the years up of a nest box and the percent success rate.

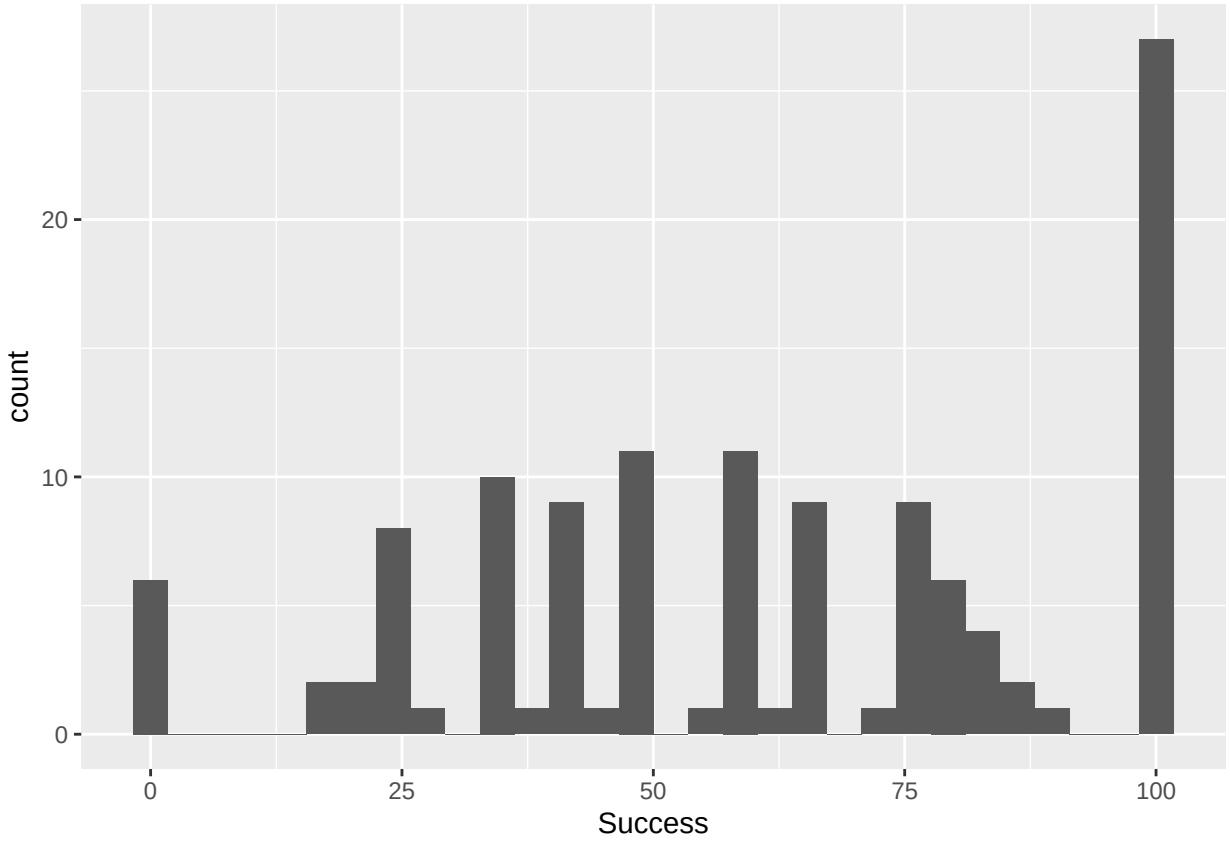
I would hypothesize that the longer the nest box is up the box will have a higher the percent success rate because there is more time for the box to house kestrels that will fledge.

Lets analyze this by making a linear regression model

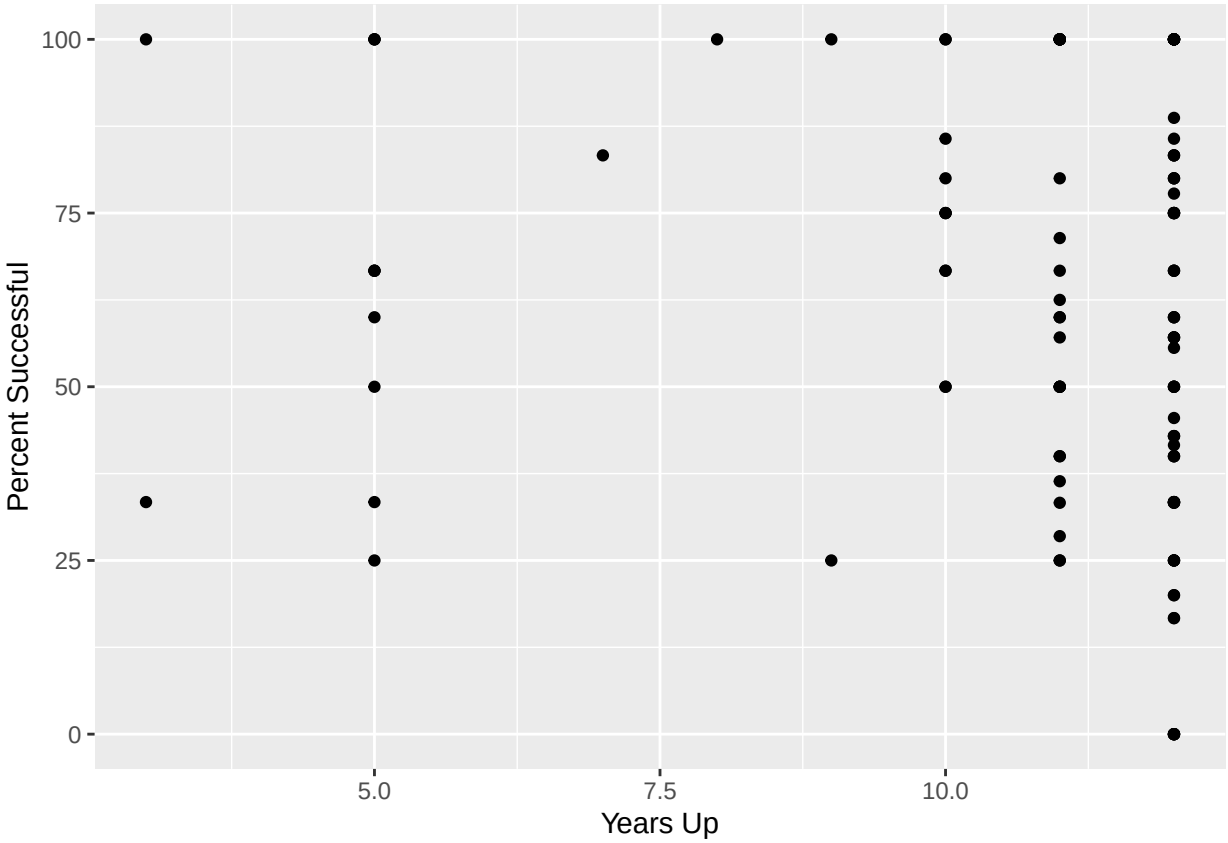
Lets make some initial plots to look at the data displayed.

```
ggplot(Kestrel_summary, aes(Success))+
  geom_histogram()
```

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



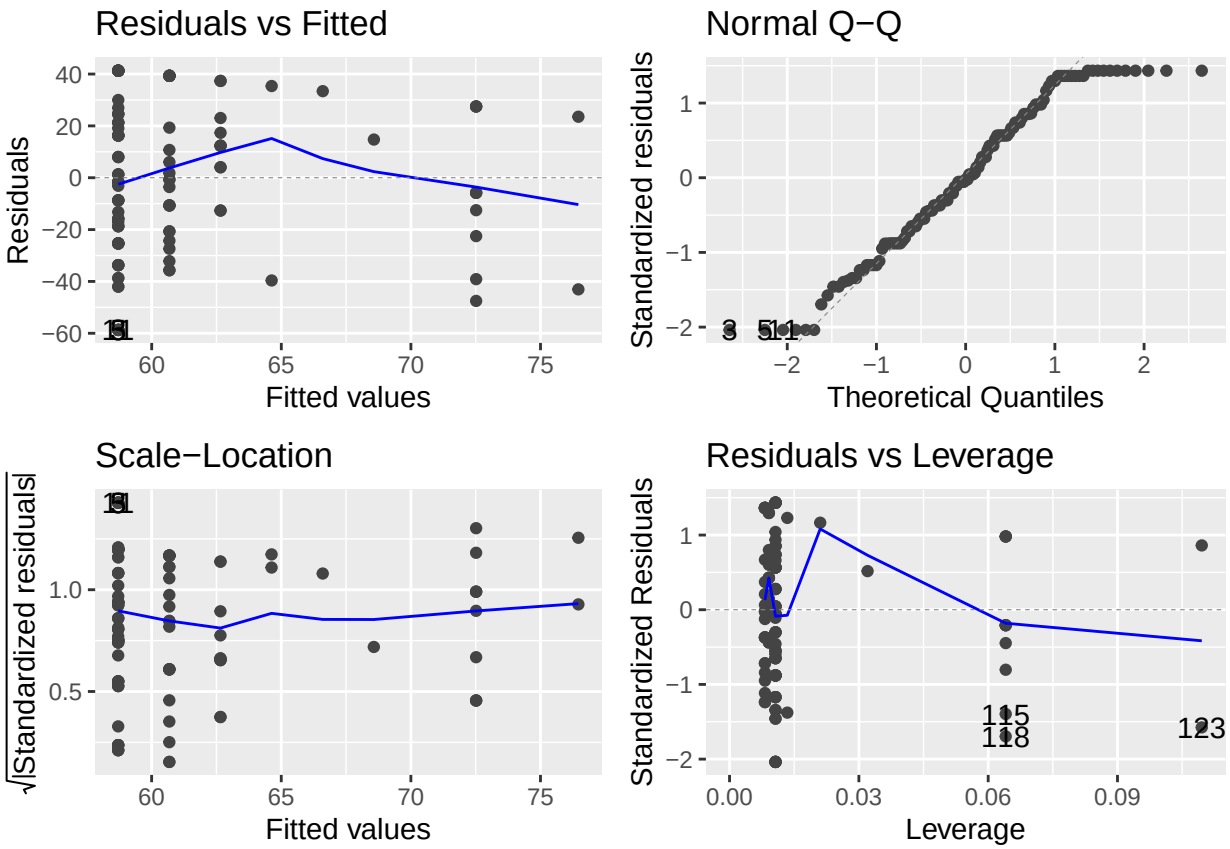

```
ggplot(Kestrel_summary, aes(Years.Up, Success)) +
  geom_point()+
  xlab("Years Up")+
  ylab("Percent Successful")
```



Based on these graphs it appears that boxes with more years up have a higher percent success rate.

Lets analyze this by making a linear regression model.

```
lmNewData <-lm(Success ~ Years.Up, data = Kestrel_summary)
autoplot(lmNewData)
```



The data looks normally distributed but with some straying at both the lower end and upper end.

View the Summary of my linear model.

```
summary(lmNewData)
```

```
##
## Call:
## lm(formula = Success ~ Years.Up, data = Kestrel_summary)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -58.710 -21.600  -0.683  24.064  41.290
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   82.379     13.048   6.314 4.69e-09 ***
## Years.Up      -1.972      1.187  -1.662   0.0991 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 28.96 on 121 degrees of freedom
## Multiple R-squared:  0.02232,    Adjusted R-squared:  0.01424
## F-statistic: 2.762 on 1 and 121 DF,  p-value: 0.0991
```

And view the ANOVA

```
anova(lmNewData)
```

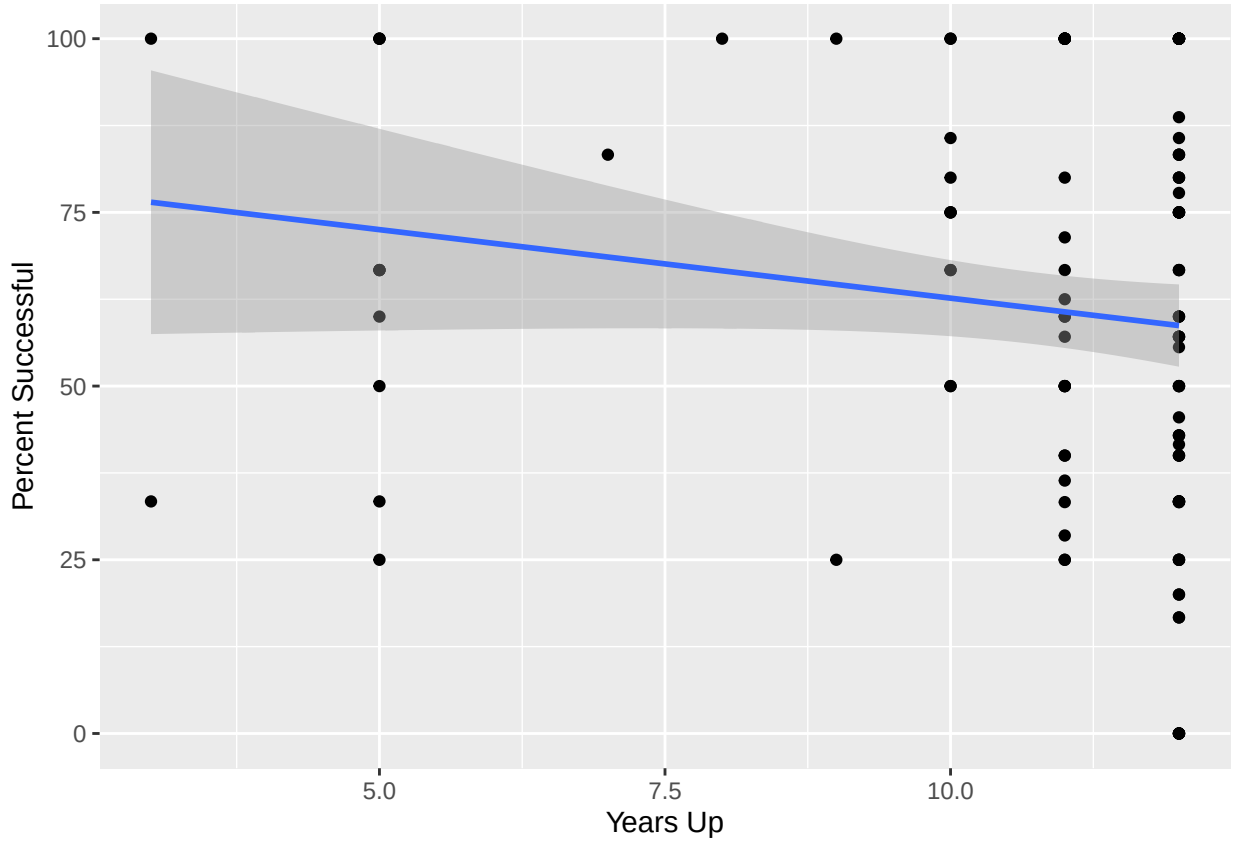
```
## Analysis of Variance Table
##
## Response: Success
##            Df Sum Sq Mean Sq F value Pr(>F)
## Years.Up    1   2317 2317.20  2.7623 0.0991 .
## Residuals 121 101502  838.86
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The results of the summary show that there is a p-value of 0.0991 meaning that there is not a significant correlation between Years Up and Success rate.

Final plot including a line of best fit

```
ggplot(Kestrel_summary, aes(Years.Up, Success))+
         geom_point()+
         geom_smooth(method = "lm")+
         xlab("Years Up")+
         ylab("Percent Successful")
```

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



This figure shows the linear regression model for the relationship between Percent Successful and Years Up. It shows that the values are not normally distributed along the linear best fit line and that there is not a significant relationship between Percent Successful and Years Up.

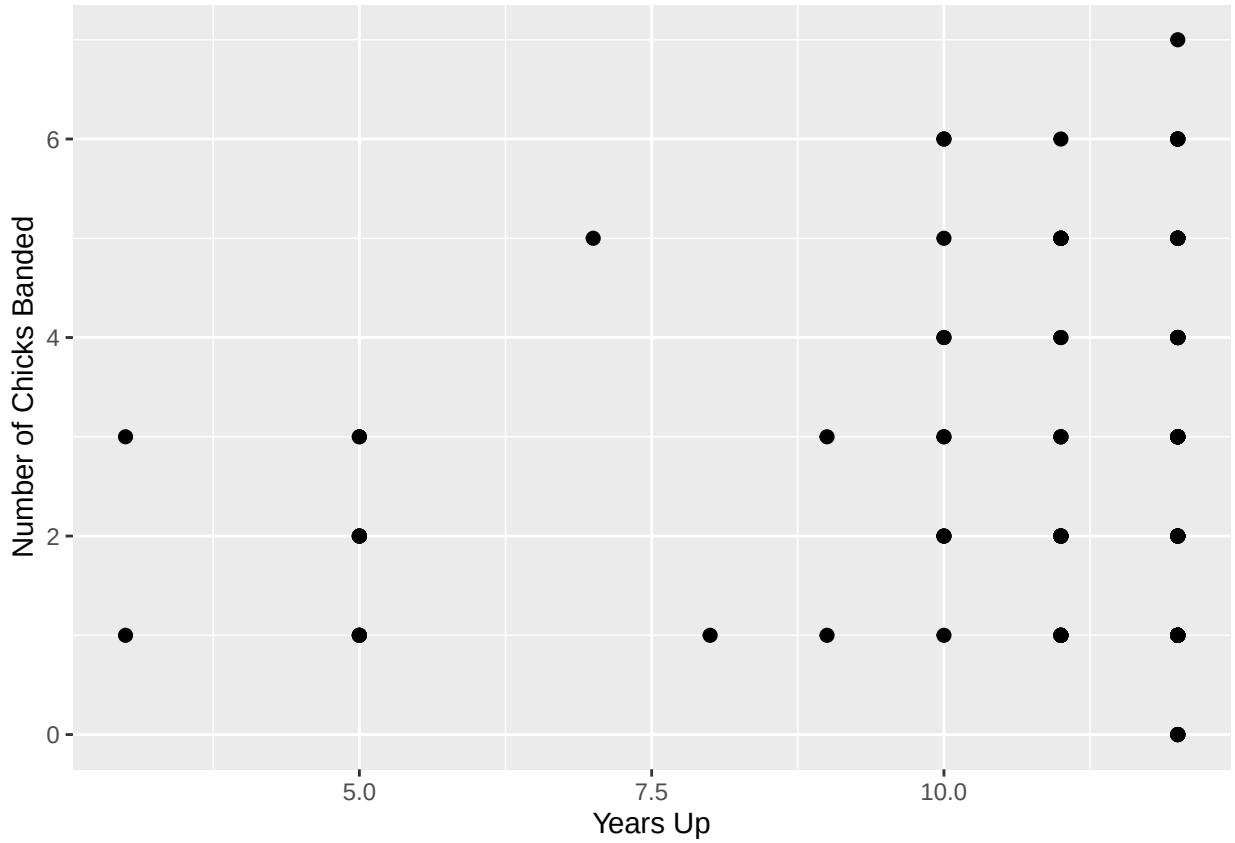
Hypothesis #2

Exploring Years Up vs Number of Chicks Banded

Does the amount of years a nest box is up impact the likelihood of chicks being present in the nest box?

Let's generate a scatter plot to look at the initial relationship between the variables.

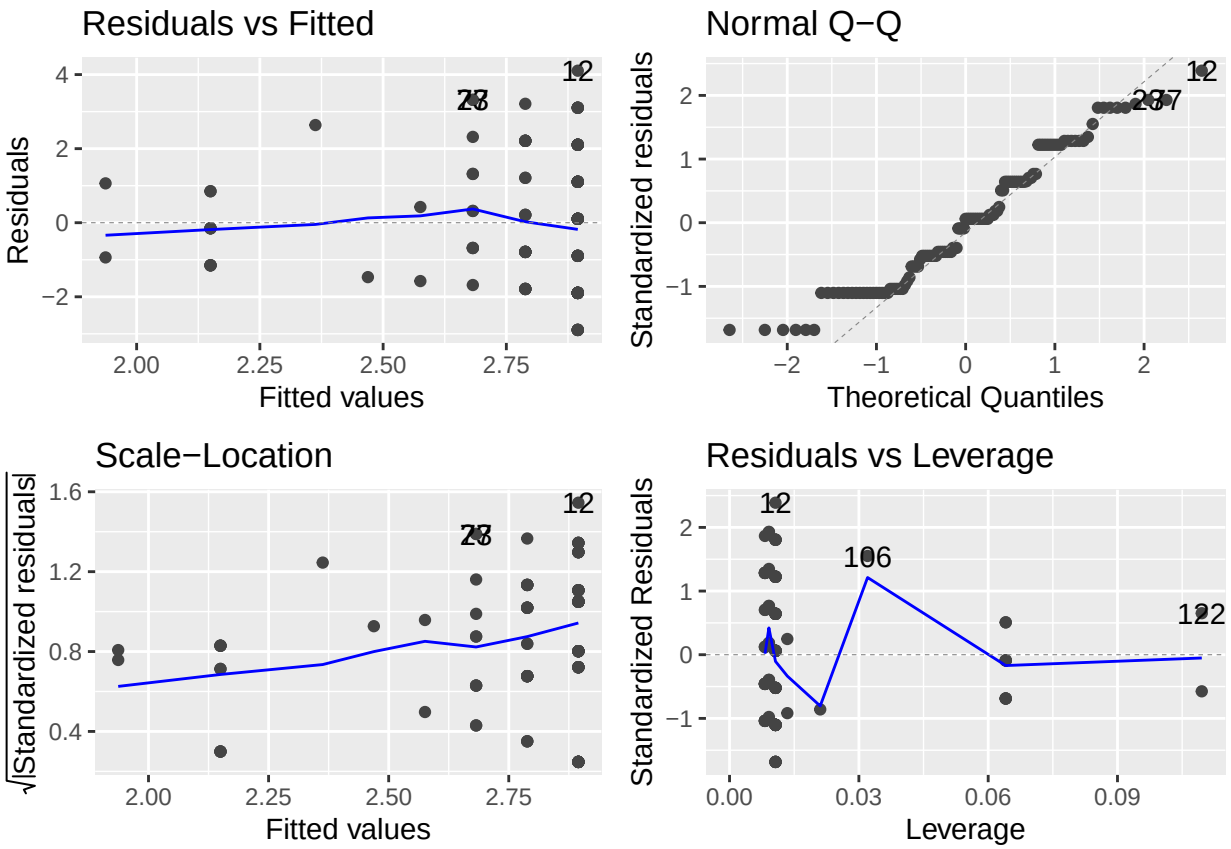
```
ggplot(Kestrel_summary, aes(x = Years.Up, y = Chicks)) + geom_point(size =2)+
xlab("Years Up") +
ylab("Number of Chicks Banded")
```



Based on this figure I would assume that the more years up a box has the more chicks banded.

For this analysis we will use the mean number of chicks banded for each amount of years up. Years Up is a categorical variable, so we will analyze this relationship using a one way anova.

```
ChickLm <-lm(Chicks ~ Years.Up, data = Kestrel_summary)
autoplot(ChickLm)
```



The data does not appear to be normally distributed, lets look the anova results and the summary results.

```
anova(ChickLm)
```

```
## Analysis of Variance Table
##
## Response: Chicks
##            Df Sum Sq Mean Sq F value Pr(>F)
## Years.Up    1   6.75  6.7481  2.2592 0.1354
## Residuals 121 361.41  2.9869
```

```
summary(ChickLm)
```

```
##
## Call:
## lm(formula = Chicks ~ Years.Up, data = Kestrel_summary)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.8949 -1.6288  0.1051  1.1051  4.1051
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.61762    0.77859   2.078   0.0399 *
## Years.Up     0.10644    0.07081   1.503   0.1354
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 1.728 on 121 degrees of freedom
## Multiple R-squared:  0.01833,    Adjusted R-squared:  0.01022
## F-statistic: 2.259 on 1 and 121 DF,  p-value: 0.1354
```

The results of this one-way anova show that there is no significant relationship between Number of Chicks Banded and Years Up, P= 0.1354.

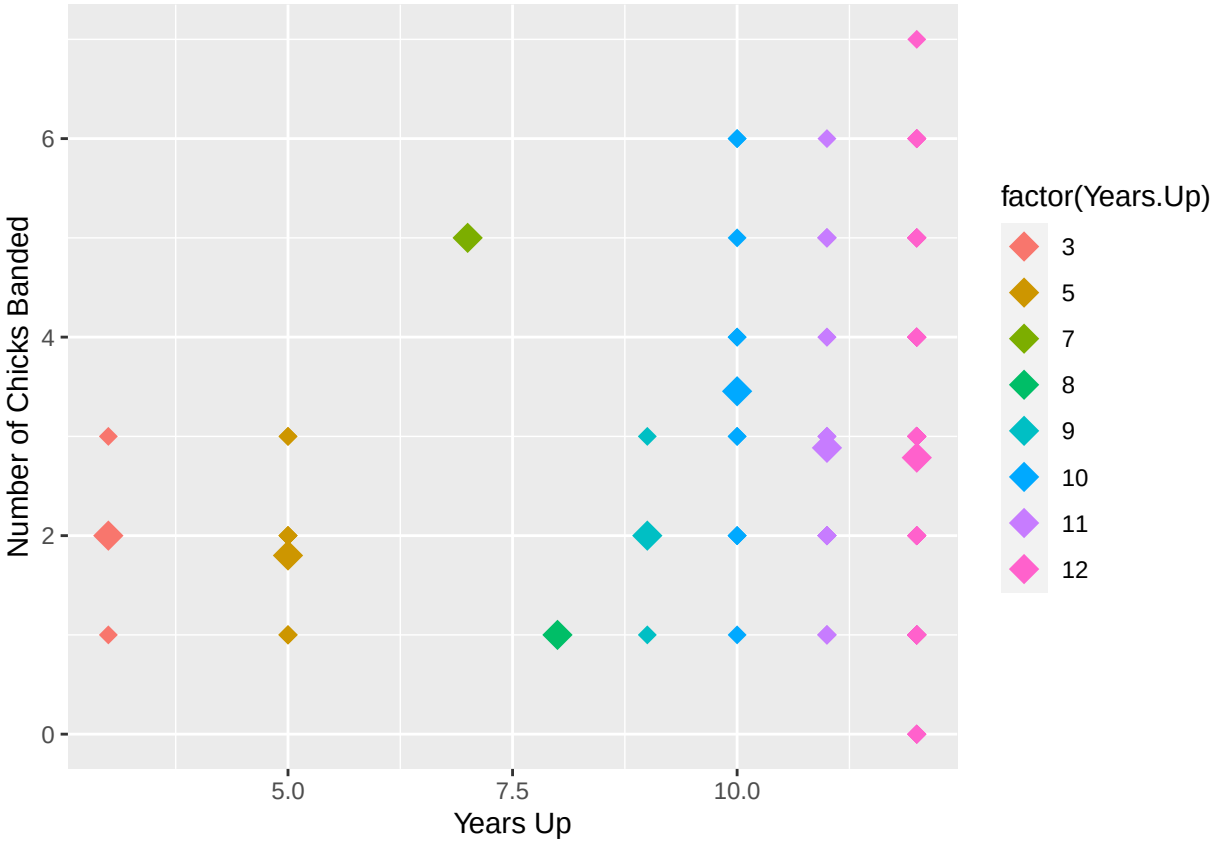
For the final plot we will need to get the mean Number of Chicks Banded for each Years Up.

```
sumDat<-Kestrel_summary %>%
  group_by(Years.Up)%>%
  summarise(Chicks = mean(Chicks))
sumDat
```

```
## # A tibble: 8 x 2
##   Years.Up Chicks
##      <int>  <dbl>
## 1        3   2
## 2        5   1.8
## 3        7   5
## 4        8   1
## 5        9   2
## 6       10   3.45
## 7       11   2.88
## 8       12   2.79
```

Then plot the data on a scatter plot while highlighting the mean chicks banded for each years up with a larger diamond point.

```
ggplot(Kestrel_summary, aes(x = Years.Up, y = Chicks, colour = factor(Years.Up)))+
  geom_point(size =3, shape = 18)+
  geom_point(data = sumDat, size = 5, shape = 18)+
xlab("Years Up") +
ylab("Number of Chicks Banded")
```



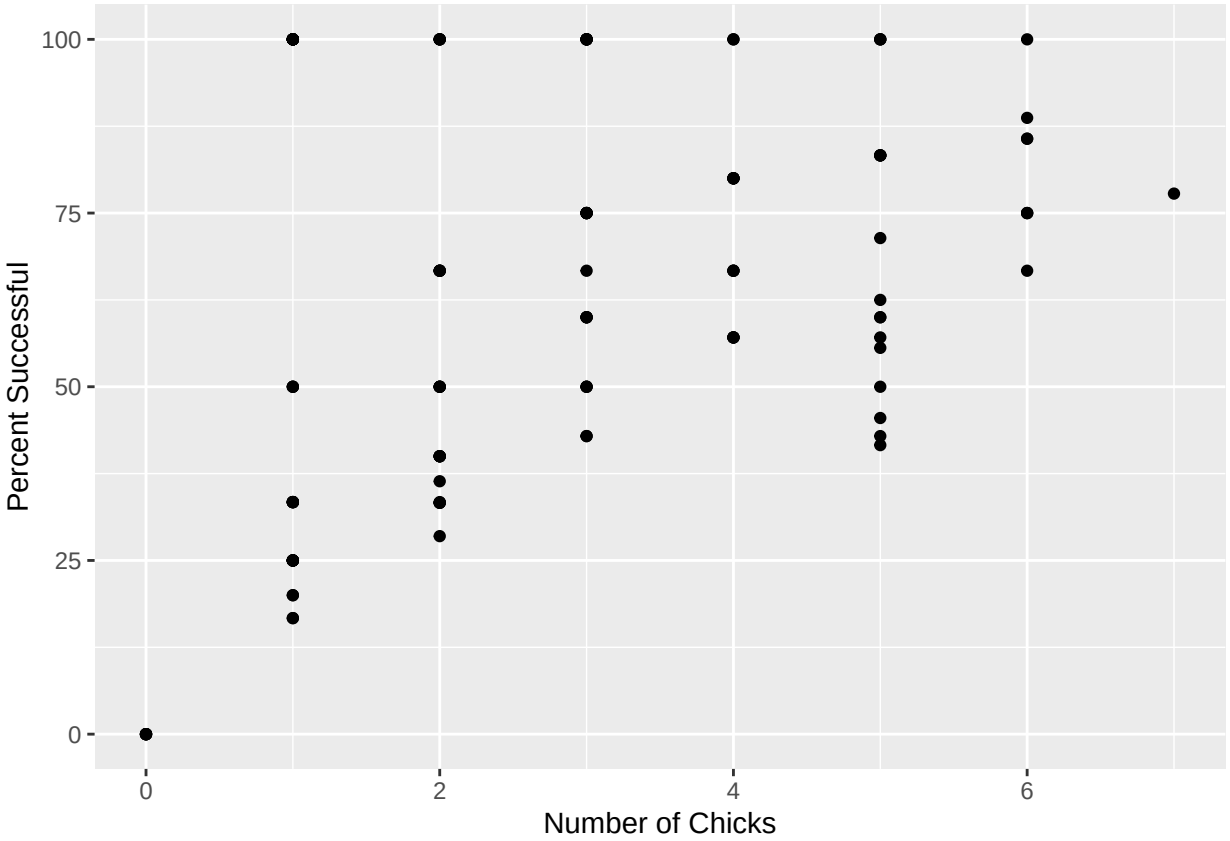
Hypothesis #3

Exploring the relationship betweek number of chicks and success rate.

Does higher number of chicks banded result in a higher success rate for the boxes?

First lets get an initial graph of the relationship

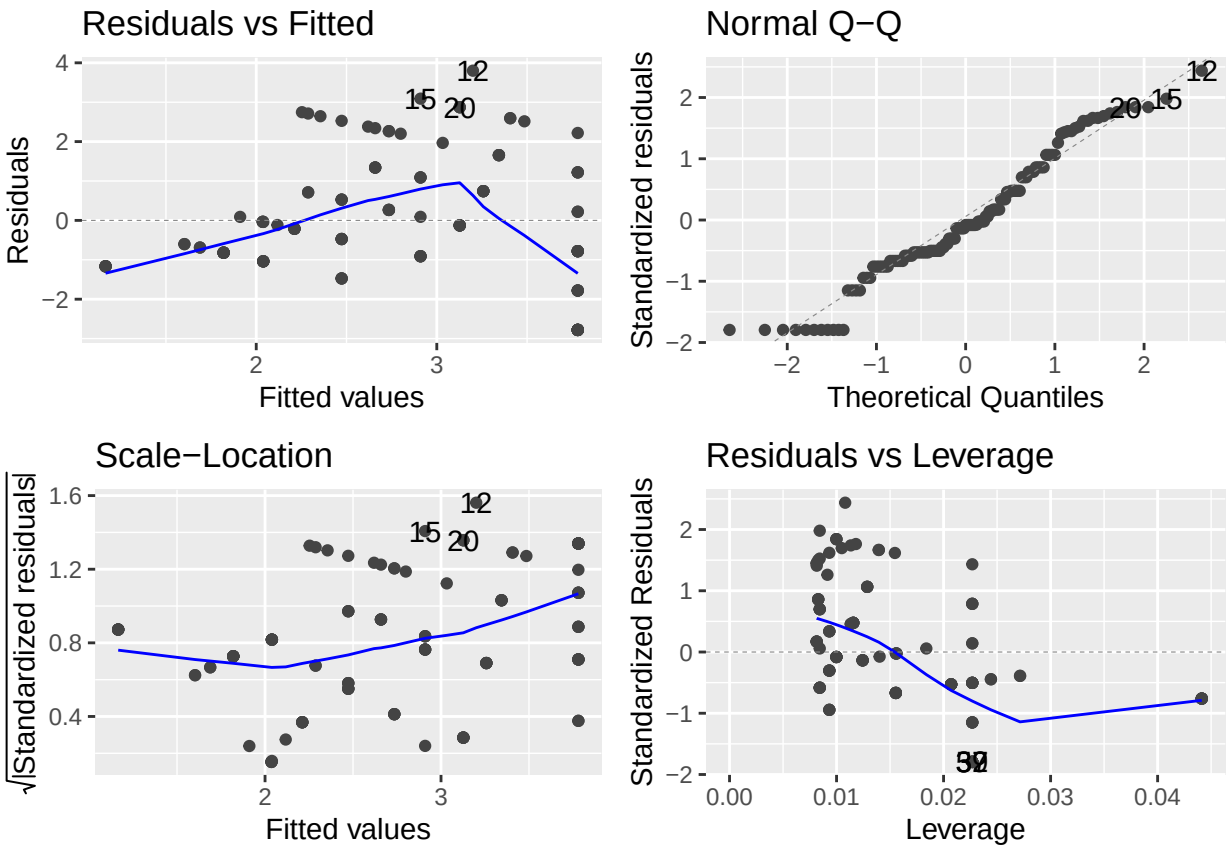
```
ggplot(Kestrel_summary, aes(Chicks, Success)) +
  geom_point()+
  ylab("Percent Successful")+
  xlab("Number of Chicks")
```



The graph appears to show that the higher the number of chicks banded, the higher the percent success rate.

We will test this using a linear regression model.

```
ChicksSuccessLm <-lm(Chicks ~ Success, data = Kestrel_summary)
autoplot(ChicksSuccessLm)
```



The data here looks to be normally distributed, however it strays at the top and towards zero.

```
anova(ChicksSuccessLm)
```

```
## Analysis of Variance Table
##
## Response: Chicks
##            Df  Sum Sq Mean Sq F value    Pr(>F)
## Success     1  70.997  70.997  28.908 3.753e-07 ***
## Residuals 121 297.166   2.456
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
summary(ChicksSuccessLm)
```

```
##
## Call:
## lm(formula = Chicks ~ Success, data = Kestrel_summary)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.7807 -0.9098 -0.1269  1.0902  3.7999
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 1.165599   0.329197   3.541 0.000567 ***
## Success     0.026151   0.004864   5.377 3.75e-07 ***
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.567 on 121 degrees of freedom
## Multiple R-squared:  0.1928, Adjusted R-squared:  0.1862
## F-statistic: 28.91 on 1 and 121 DF,  p-value: 3.753e-07
```

The results of this one-way anova show that there is a significant relationship between Number of Chicks Banded and Percent Success Rate, P= 3.753e-07. This means that we can reject the null hypothesis that there is no correlation between Number of Chicks Banded and Percent Success Rate.

Lets make a final plot of this. First we will need to find the mean success rate for each amount of chicks banded.

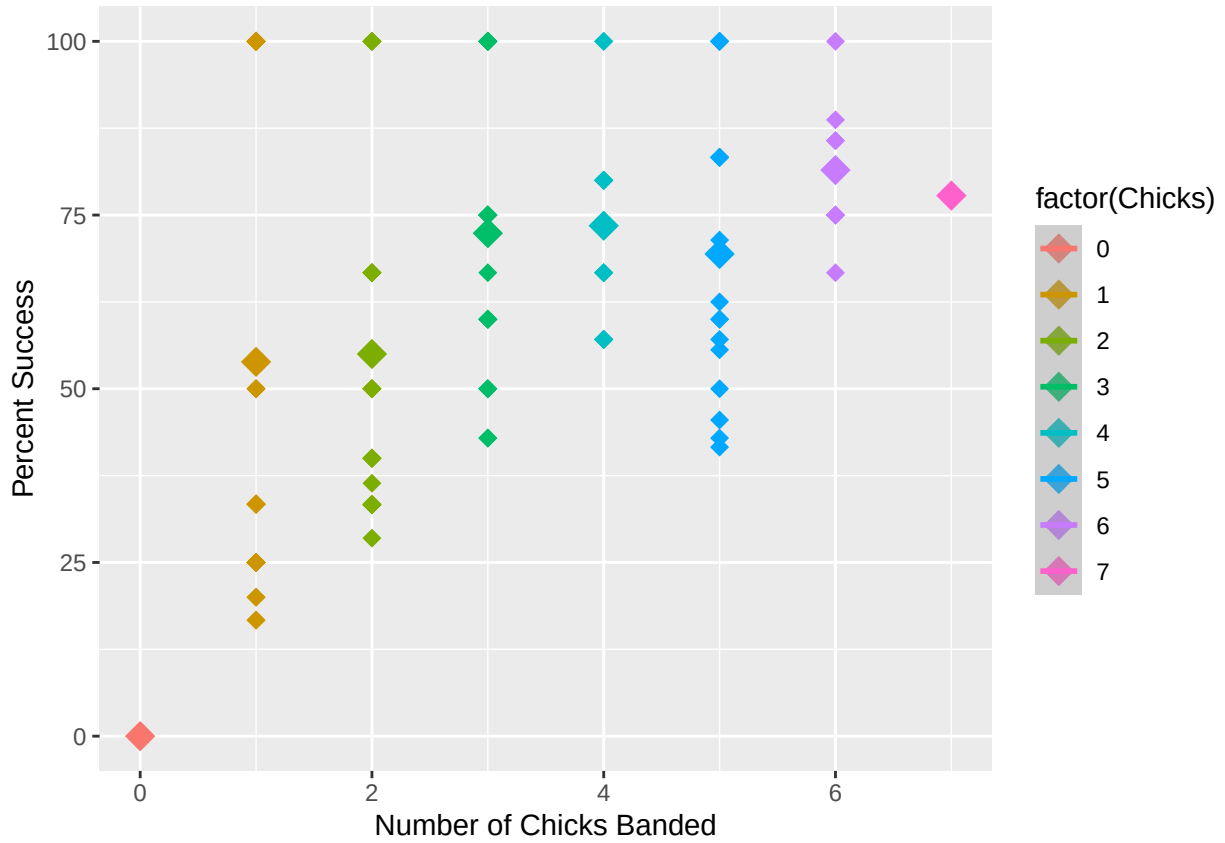
```
sumSuccess<-Kestrel_summary %>%
  group_by(Chicks)%>%
  summarise(Success = mean(Success))
sumSuccess
```

```
## # A tibble: 8 x 2
##   Chicks Success
##    <int>   <dbl>
## 1      0     0
## 2      1    53.9
## 3      2    55
## 4      3    72.4
## 5      4    73.5
## 6      5    69.4
## 7      6    81.5
## 8      7    77.8
```

Then we will plot the data and add a layer that will plot the mean success rates over top as a larger diamond.

```
ggplot(Kestrel_summary, aes(x = Chicks, y = Success, colour = factor(Chicks)))+
  geom_point(size =3, shape = 18)+
  geom_point(data = sumSuccess, size = 5, shape = 18)+
  geom_smooth(method = "lm")+
xlab("Number of Chicks Banded") +
ylab("Percent Success")
```

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



This graph shows the steady increase in nest box percent success as number of chicks banded increases. The different amount of chicks banded are each uniquely colored and the mean success rate for each number of chicks is highlighted as a larger diamond. This helps to visualize the positive relationship between percent success and number of chicks banded.

Biological Summary

Overall, what I found in this analysis is that there is no significant relationship between number of years up and nest box percent success, or between number of years up and the amount of chicks being present in the nest box. This analysis did however find that there is a significant relationship between number of chicks banded and percent success.

My first hypothesis was that boxes with more years up would have a higher nest box success rate for kestrels because there would be a longer period of time available for kestrels to nest in the box. The analysis rejected this hypothesis because there was a p-value of 0.0991. This means that there was no significance between nest box success rate and years up, so we can accept the null hypothesis.

My second hypothesis was that boxes with more years up would yield a higher number of chicks banded because there would be more time to for kestrels find and recognize the nest box as a nesting option. My analysis rejected this hypothesis because there was a p-value of 0.1354. This means that there was no significance between number of chicks banded and number of years up a box has, so we can accept the null hypothesis.

My third and final hypothesis was that nest boxes with a higher number of chicks banded would result in a higher success rate for the boxes because if more chicks are banded they have a higher chance of fledging and being successful. My analysis proved that my hypothesis was correct because there was a p value of

3.753e-07. So we can reject the null that there is no significant relationship between number of chicks banded and percent success rate.

Challenges

I am not very familiar with statistics so the statistical testing portion of the CapDap was definitely challenging for me. Specifically determining which test would be best for my relationships. I had to reread the Getting Started with R book over many times to get an idea of where to go. When that did not help me I would use the internet or ask for help from fellow students in the PQRC.

Another challenge I encountered was how to display my data in the best way. How I overcame this challenge was again, rereading the textbook and consulting the internet for help.

Lastly, one of my major challenges I found was working with my data set. The data set I chose to work with was very untidy and required a lot of fixing up before I could start any of my analysis. This meant I spent a lot of time reading the book, and using the internet to find different ways to help me adjust my data so that I could effectively plot my relationships and run my tests.

What I learned from my challenges is that the way you record data is very important. It is important that what you enter is uniform so that there are less issues in R and so that whoever is analyzing it does not have to spend a lot of time tidying up the data. I also learned new functions like “if else” that I did not end up using in my final, but was still excited to learn.