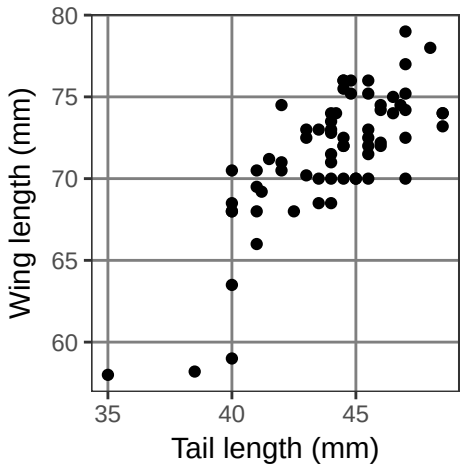
Using the words we learned in this class, answer all questions succinctly; you will lose points for rambling. If you provide R code to answer a problem, provide all the necessary code and indicate clearly which variable(s) contain the answer.

1. Below is a boxplot of infection risk by region from a random sample of hospitals from around the United States. Provide the R code to run the **correct** ANOVA of these data.

```
hospitals %>%
    ggplot(aes(factor(region), infection_risk)) + geom_boxplot() +
    labs(x="Region", y="Infection risk")

fit <- lm(infection_risk ~ factor(region), data=hospitals)
anova(fit)
```

2. Consider the variables `taillength` and `winglength`, below, for finches from the Galapagos islands.

```
finches <- read.csv("/Users/ez/website/app/public/data/finches.csv")

ggplot(finches, aes(taillength, winglength)) +
    geom_point() +
    labs(x="Tail length (mm)", y="Wing length (mm)")
```



   (a) Provide as much R code as you can to mimic the plot above.

   (b) Write R code to make an appropriate 93% confidence interval for these data.

```
x <- with(finches, taillength - winglength)
mean(x) + qt(c(0.035, 0.965), length(x) - 1)*sd(x)/(length(x) - 1)

## [1] -27.61988 -27.47424
```

(c) Set up the appropriately matching hypotheses for the confidence interval above.
$H_0 : \mu_d = 0$ versus $H_1 : \mu_d \neq 0$ at $\alpha = 0.07$

(d) Suppose you computed the interval $(-30.02, -28.04)$. Interpret the provided confidence interval, clearly stating which variable's mean is larger, if any.
We are 93% confident that the true mean difference between winglength and taillength is between -30.02 and -28.04 milimeters, when subtracting winglength from taillength. Hence, the mean for wing length is larger.

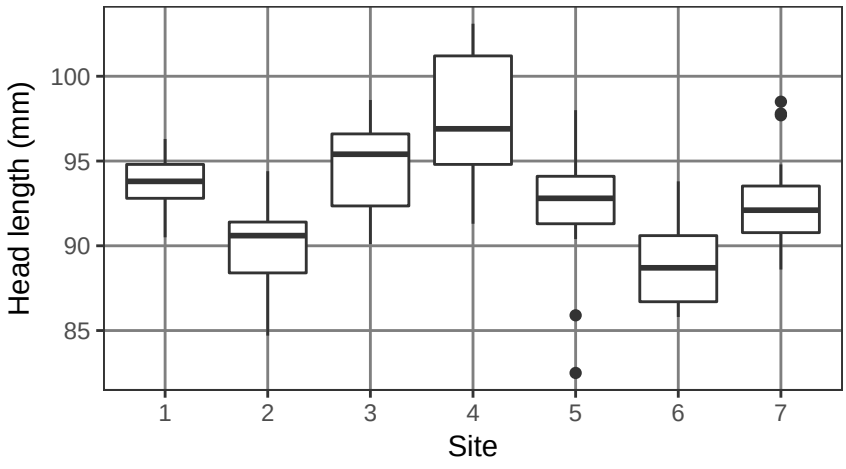
(e) Does the confidence interval provide convincing evidence that there is a real difference in the mean wing to tail length? Explain.
Because the confidence interval does not include 0, we can say there is sufficient evidence of a real difference.

(f) Explain to my grandmother the statistical conclusion from your confidence interval.
Yo, grams, wings are on average longer than tail for finches from the galapagos.

(g) Would your conclusion differ if instead a 99% confidence interval was used? Explain. Maybe, as a 99% confidence interval would be wider. Though, without doing the calculations we can't say for sure if 0 would be included in the new interval.

3. Consider a random sample of opossum from multiple sites within Australia. Head length was measured in milimeters.

```
possum <- read.csv("/Users/ez/website/app/public/data/possum.csv")
possum$site <- factor(possum$site)

possum %>%
    ggplot(aes(site, headL)) +
    geom_boxplot() +
    labs(x="Site", y="Head length (mm)")
```



```
model <- lm(headL~site, data=possum)
anova(model)

## Analysis of Variance Table
##
## Response: headL
##           Df Sum Sq Mean Sq F value    Pr(>F)
## site       6 499.94  83.323   9.914 1.629e-08 ***
## Residuals 97 815.25   8.405
## ---
## Signif. codes:
## 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

(a) Provide as much R code as you can to mimic the plot above.

(b) Set up an appropriate hypothesis test for the analysis above. $H_0 : \mu_1 = \ldots = \mu_7$ versus H_1 : at least one mean is different at $\alpha = 0.01$.

(c) Conclude the hypothesis test. Because p-value $< 0.0001 < \alpha = 0.01$, we reject H_0 in favor of the alternative. At least one mean is different from the rest.

(d) Are the assumptions of ANOVA satisfied. Explain each with a complete, English sentence. Normality is met since there is very little skew. The constant variation is met since there is about equal variation within each group. We collected a random sample so that helps ensure independent data.

(e) Provide the R code to appropriately test the two groups' means that you believe are most different.

```
possum %>%
    filter(site %in% c(2, 4)) %>%
    t.test(headL ~ site, data=., conf=.99)

##
##  Welch Two Sample t-test
##
## data:  headL by site
## t = -4.2117, df = 9.3443, p-value = 0.002086
## alternative hypothesis: true difference in means is not equal to 0
## 99 percent confidence interval:
##  -13.89669  -1.85496
## sample estimates:
## mean in group 2 mean in group 4
##        89.73846        97.61429
```

4. Suppose a new graduate is considering a job in two locations, Cleveland, OH and Sacramento, CA, and she wants to see whether the average income in one of these cities is higher than the other. She runs the following R code.

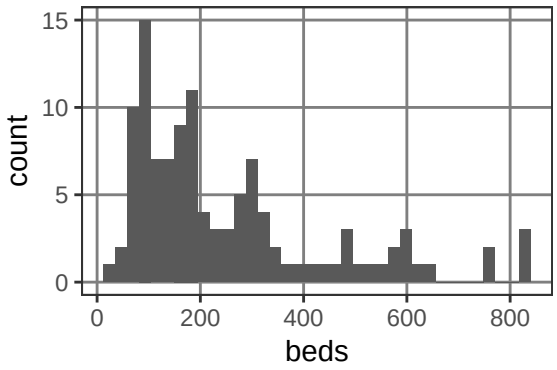
```
t.test(income~city, data=city_incomes, conf.level=0.98)


##
##  Welch Two Sample t-test
##
## data:  income by city
## t = -2.5744, df = 101.68, p-value = 0.01148
## alternative hypothesis: true difference in means is not equal to 0
## 98 percent confidence interval:
##  -26314.902  -1123.439
## sample estimates:
##  mean in group Cleveland mean in group Sacramento
##                 20410.18                 34129.35
```

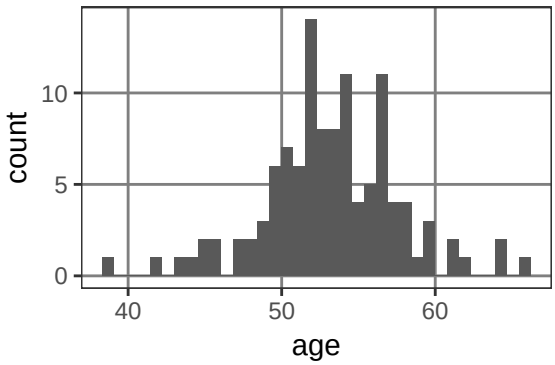
(a) Set up, evaluate, and conclude the hypothesis test implicit to the `R` code above. $H_0 : \mu_C = \mu_S$ versus $H_1 : \mu_C \neq \mu_S$ at $\alpha = 0.02$. Because p-value $= 0.015 < \alpha = 0.02$, we reject H_0 and conclude that Sacramento's mean income is higher than Cleveland's.

(b) Does the provided confidence interval agree with your conclusion? Explain. Yes, since the confidence interval excludes 0.

(c) Explain the literal meaning of a $P\%$ confidence interval. 98% of the theoretically resampled confidence intervals will contain the true difference in mean incomes between Sacramento and Cleveland.

5. Consider a random of sample of hospitals from the United States, where characteristics of both the hospital and the patients within each hospital were collected. The variable `beds` records the number of beds in the hospital. The variable `age` records average patient age of patients within each hospital.





(a) Describe the shape of data for each variable above. Beds is right skewed and age is fairly symmetric.

(b) I, Edward, claim that the shape of these data, given their descriptions, was expected. Explain why. Because there are inherently some large hospitals, but most are small, we assume that beds will be right skewed and it is. On the other hand, age is a mean within each hospital hence it should be symmetric based on the CLT and it is.