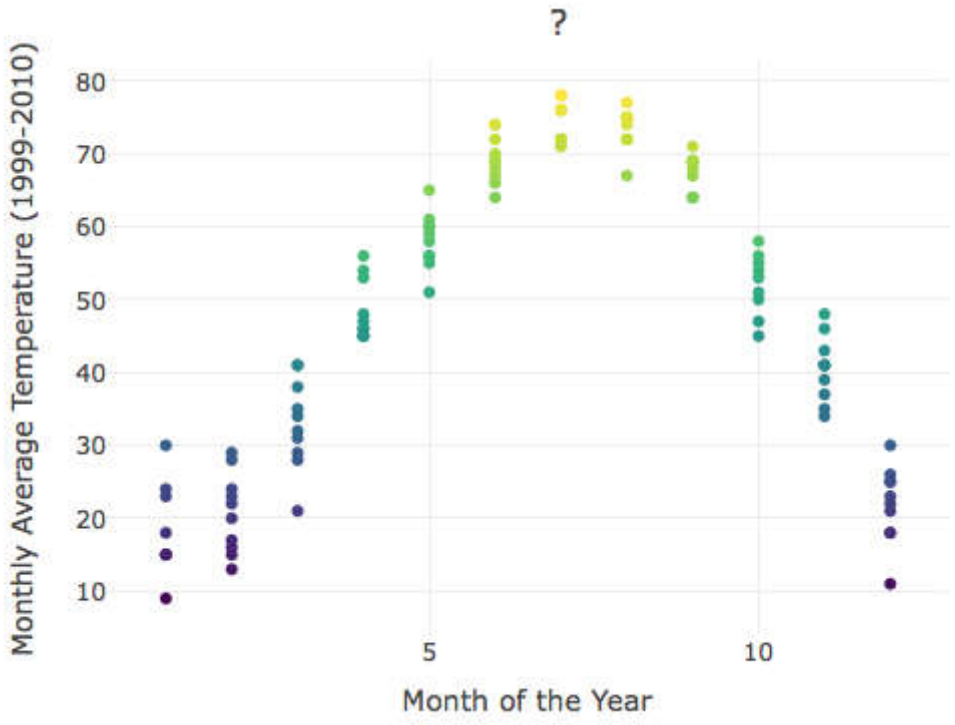
[PRINT]

Math 325 - Fall 2020 - Section 2,
Kyle Tolliver, 12/13/20 at 12:45:06 AM MST

Question1: *Score 1/1*

Which of the following would provide the most meaningful title to this graphic?



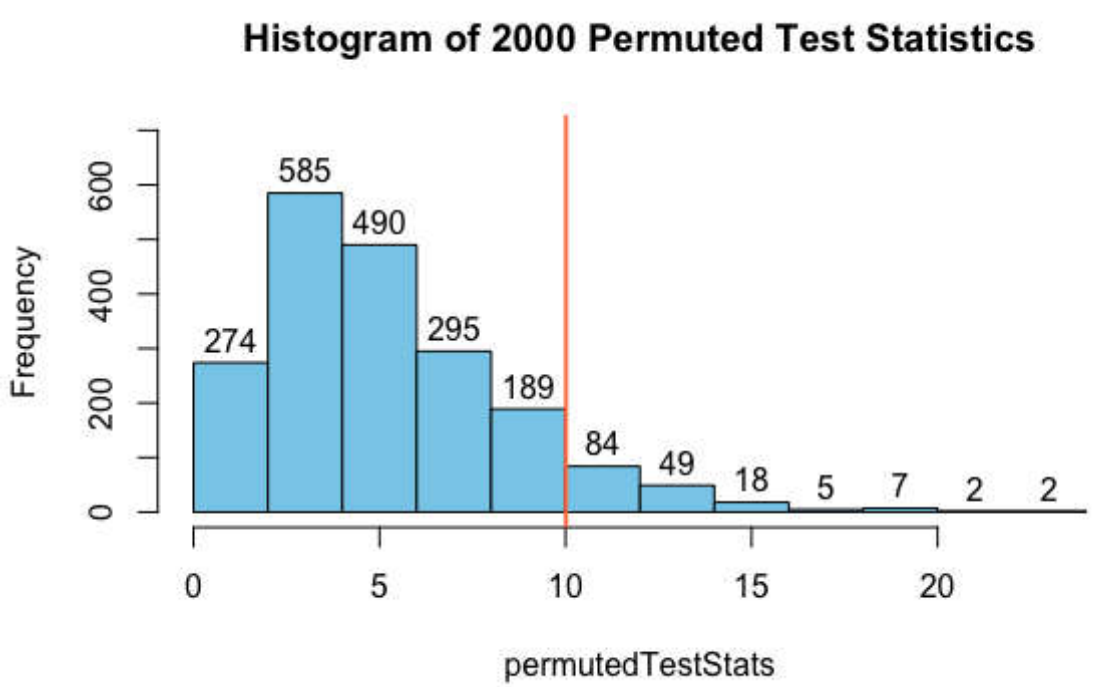
Your response	Correct response
Beautiful Summers but Freezing Winters	

Auto graded **Grade:** 1/1.0

Total grade: 1.0×1/1 = 100%

Question2: *Score 1/1*

A permutation test for a Kruskal-Wallis Test Statistic is performed with the following result:



Note that the numbers above the bars state the frequency of observations contained in that bar.

What is the p-value of this permutation test?

Your response	Correct response
p-value = 0.0835	

Auto graded **Grade:** 1/1.0

Total grade: 1.0×1/1 = 100%

Feedback:

The p-value is found as the percentage of permuted test statistics that are more extreme than the observed test statistic (the red line). In this case, the percentage more is found by adding up the number of observations to the right of the red line: 84 + 49 + 18 + 5 + 7 + 2 + 2 = 167 and 167/2000 = 0.0835. We divide by 2000 because that is how many test statistics there were (see the main title of the graph). Note that we do not double this p-value because the test statistic is always positive and right skewed for the Kruskal-Wallis Test.

Question3: *Score 1/1*

Run the following code in R.

```
Util <- Utilities
Util$Season <- cut(Util$month%%12, c(0,2,5,8,11), c("Winter","Spring","Summer","Fall"))
boxplot(elecbill ~ Season, data=Util, main="A Residence in Minnesota", ylab="Monthly Electricity Bill (US Dollars)", xlab="Season of the Year", col=c("skyblue","darkseagreen3","coral","goldenrod"))
```

Perform an appropriate test in R to determine if at least one season shows evidence of yielding stochastically different values of electric bills from the other seasons.

Report the test statistic, degrees of freedom, and p-value of the test.

Your response	Correct response
chi-squared = 18.677, df = 3, p-value = 0.0003189	

Auto graded Grade: 1/1.0

Total grade: 1.0×1/1 = 100%

Feedback:

Because the distributions don't all have the same shape (and aren't all normal) a Kruskal-Wallis Test is more appropriate than an ANOVA:

```
> kruskal.test(elecbill ~ Season, data=Util)

    Kruskal-Wallis rank sum test

data:  elecbill by Season
Kruskal-Wallis chi-squared = 18.677, df = 3, p-value = 0.0003189
```

Question4: *Score 1/1*

What does each row of the Utilities dataset represent?

Your response	Correct response
A billing period.	

Auto graded Grade: 1/1.0

Total grade: 1.0×1/1 = 100%

Feedback:

Running the codes

?Utilities

View(Utilities)

reveals that each row corresponds to a billing period. The columns of the data set give different characteristics or measurements on that billing period including the month it was in, the day of the month it ended on, the price of various bills, the temperature and so on. But the "observations" or the "rows" of the dataset are the various billing periods this particular homeowner experienced.

Question5: *Score 1/1*

Run the following codes in R.

```
util.lm <- lm(gasbill ~ temp + I(temp^2), data=Utilities)
plot(gasbill ~ temp, data=Utilities)
b <- util.lm$coefficients
curve(b[1] + b[2]*x + b[3]*x^2, add=TRUE)
summary(util.lm)
```

Which of the following correctly states the mathematical model for the above regression?

Your response	Correct response
$\underbrace{Y_i}_{gasbill} = \beta_0 + \beta_1 \underbrace{X_i}_{temp} + \beta_2 X_i^2 + \epsilon_i$	

Auto graded Grade: 1/1.0

Total grade: 1.0×1/1 = 100%

Feedback:

This regression is technically beyond the scope of this course. It applies what is called a polynomial regression model because it uses x and x^2. This is what places the curved line on the graph. It is a very useful type of regression though, and it should be within your ability to recognize the mathematical model from R code, even if that model is unfamiliar to you. Notice how X_i is used twice in the model because temp and I(temp^2) are used in the lm(...).

$$\underbrace{Y_i}_{gasbill} = \beta_0 + \beta_1 \underbrace{X_i}_{temp} + \beta_2 \underbrace{X_i^2}_{temp^2} + \epsilon_i$$

Question6: *Score 1/1*

Run the following codes in R.

```
U2 <- subset(Utilities, month %in% c(3,6))
plot(gasbill ~ elecbill, data=U2, col=month)
```

Perform an analysis that allows you to place two separate lines on this scatterplot.

Report the p-value of the test of whether these lines have significantly different slopes or not.

Your response	Correct response
No significant difference in slopes, p-value = 0.8883.	

Auto graded Grade: 1/1.0

Total grade: 1.0×1/1 = 100%

Feedback:

To allow the slopes to differ so that they can be tested for equality, an interaction term must be used:

```
> mylm <- lm(gasbill ~ elecbill + month + elecbill:month, data=U2)
> summary(mylm)
```

```
Call:
lm(formula = gasbill ~ elecbill + month + elecbill:month, data = U2)

Residuals:
    Min      1Q  Median      3Q     Max
-99.423  -2.502   0.874   4.819  60.851

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)   211.593595  77.465638   2.731   0.0148 *
elecbill       -0.008898   1.119290  -0.008   0.9938
month         -33.601545  15.341555  -2.190   0.0437 *
elecbill:month  0.031648   0.221808   0.143   0.8883
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 34 on 16 degrees of freedom
Multiple R-squared:  0.71,    Adjusted R-squared:  0.6557
F-statistic: 13.06 on 3 and 16 DF,  p-value: 0.0001437
```

Question7: *Score 1/1*

A group of Math 221 students performed a reseearch project for their class studying the parking abilities of men and women at BYU-Idaho. Essentially, they watched people park and recorded the final parking as "well done", "crooked", or "failed". They also recorded the gender of the person parking the vehicle.

What type of analyses would best help them determine if men or women are better at parking?

Your response	Correct response
Chi Squared Test	

Auto graded Grade: 1/1.0

Total grade: 1.0×1/1 = 100%

Feedback:

Because they have recorded two qualitative variables (gender and parking ability) the only appropriate analysis from the options listed would be a chi squared test. This would allow them to determine if gender and parking ability were independent or not.

Question8: *Score 1/1*

What is the correct conclusion to make in a logistic regression when the goodness of fit test (either one) gives a p-value of 0.352?

Your response	Correct response
Continue to believe the logistic regression was appropriate to perform on the given data.	

Auto graded Grade: 1/1.0

Total grade: 1.0×1/1 = 100%

Question9: *Score 1/1*

Suppose the homeowner of the home where the Utilities dataset was collected from approached you with the question: "How much should I plan on my electric bill costing me this July?"

Use the Utilities dataset in R to perform some useful calculations in order to answer their question.

Then, select the best answer from the list below.

Your response	Correct response
Based on your past data, it looks like you have spent as little as $29.32 and as much as $114.90 on electricity in July. However, 3 out of every 4 bills in July was between $91.55 abd $114.90. So planning on at least $90 is probably a safe bet.	

Auto graded Grade: 1/1.0

Total grade: 1.0×1/1 = 100%

Feedback:

The low end of electric bills in July at $29.32 seems to be an outlier and happened when the house was "empty" as witness by the "notes" section of the Utilities dataset. The first quartile of electric bills is $91.56, thus 75% of bills (or 3 out of 4) are above that value. This shows that it is rare that a bill is less than that amount, and far more likely that it would be above that amount.

Question10: *Score 1/1*

A common mistake of introductory statistics students is to make a graph like the following.

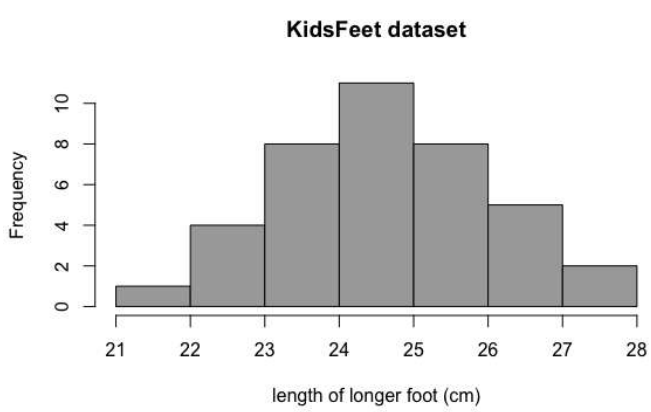
```
barplot(KidsFeet$length)
```

Consider this graph and the KidsFeet dataset yourself by running these codes in R:

```
barplot(KidsFeet$length)
View(KidsFeet)
```

Select the graph from the options below that would be more appropriate for visualizing the distribution of this data.

Your response	Correct response



Auto graded **Grade:** 1/1.0

Total grade: 1.0×1/1 = 100%

Feedback:

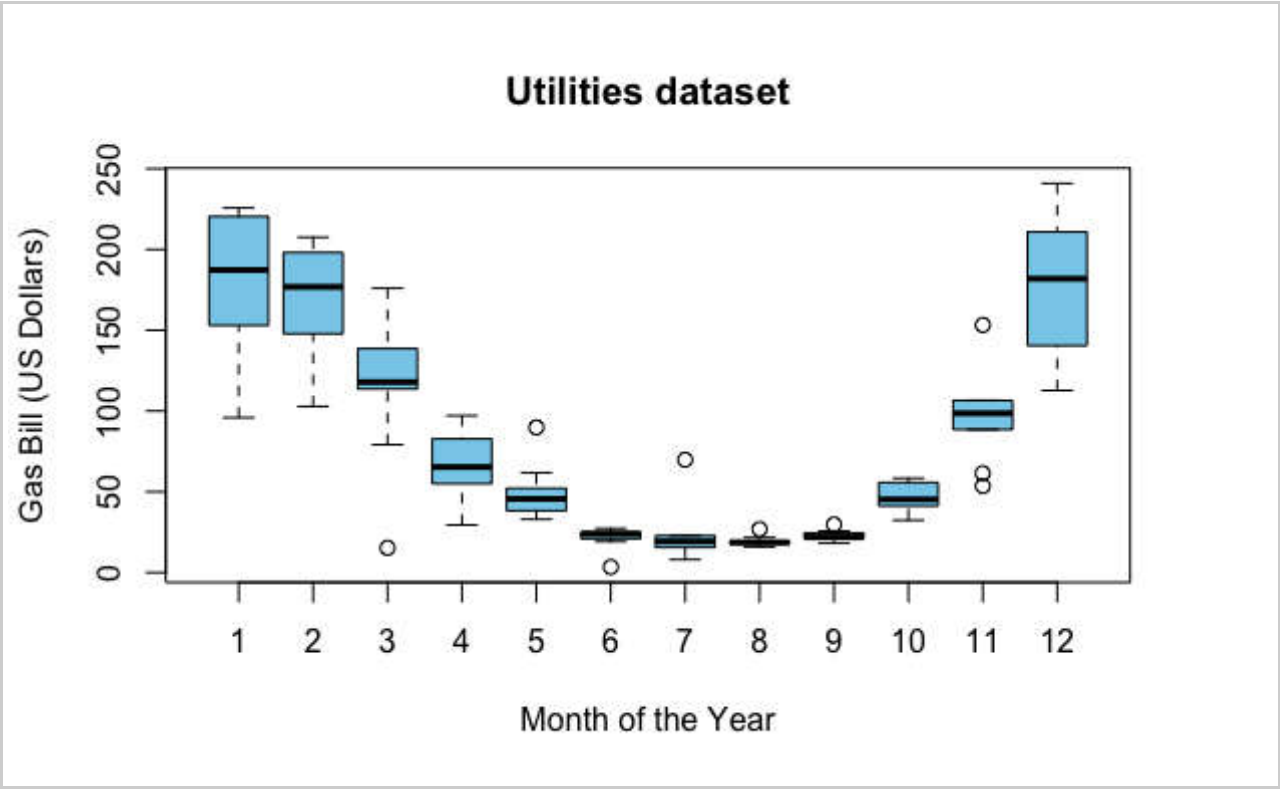
It is common that students try to depict each value instead of summarizing the overall shape of the distribution of values. However, the best way to learn from data is to summarize the overall pattern, like with a histogram in this case since "length" is a single quantitative variable.

Question11: *Score 1/1*

Which type of analysis would be best for analyzing the data shown in this boxplot?

Your response	Correct response
A Kruskal-Wallis Test	

Auto graded **Grade:** 1/1.0



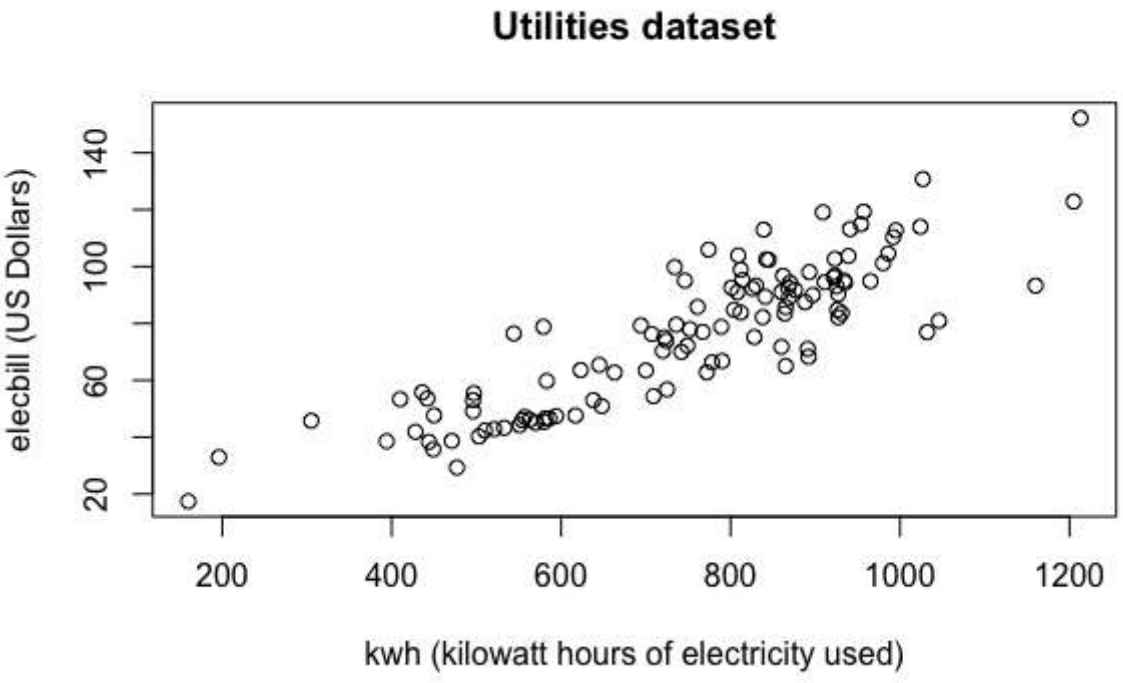
Total grade: 1.0×1/1 = 100%
Feedback:

Because the variance seems to differ dramatically between the various boxes in the boxplot, the Kruskal-Wallis Test would be better than ANOVA. This is because ANOVA requires the groups of data to have equal or "constant" variance.

Question12: *Score 1/1*

Use the Utilities dataset in R to perform a linear regression that would place a simple linear regression line on this graph.

Which of the following correctly interprets the slope of the line from this regression? (You may need to scroll down to see the answer options.)



Your response	Correct response
This homeowner is being charged an average of 10.9 cents per kilowatt hour of electricity used.	

Auto graded Grade: 1/1.0

Total grade: 1.0×1/1 = 100%

Feedback:

summary(lm(elecbill ~ kwh, data=Utilities))

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	-5.000714	4.521874	-1.106	0.271	
kwh	0.108754	0.005816	18.698	<2e-16	***

And remember, the regression slope tells us about the change in the average bill amount, not the actual bill amount.

Question13: *Score 1/1*

Consider the output of a chi squared test performed in R that is shown below. What is the p-value of the test?

Pearson's Chi-squared test

data: table(gender, groupId))

X-squared = 4.9575, df = 2, p-value = _____

Your response	Correct response
0.0838	

Auto graded Grade: 1/1.0

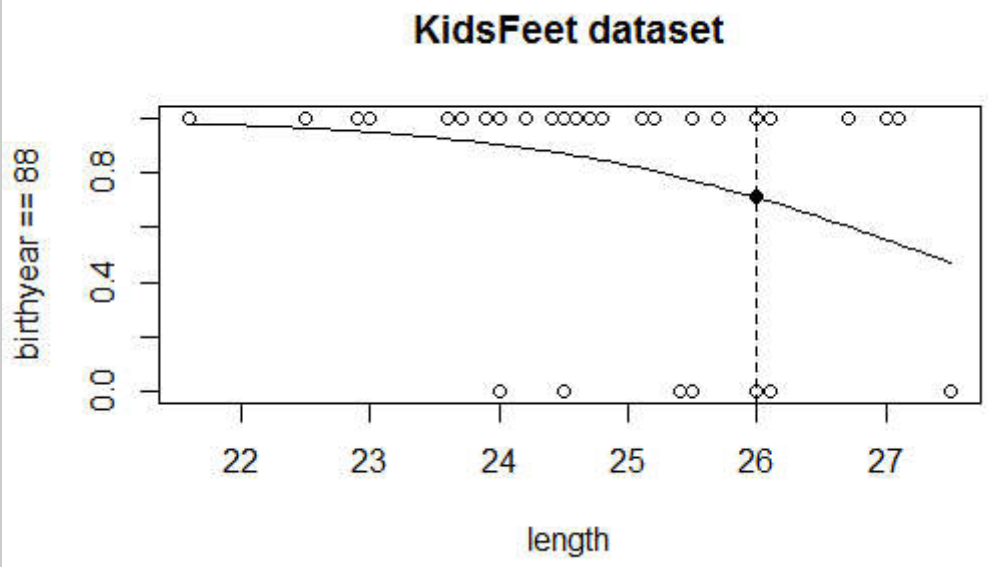
Total grade: 1.0×1/1 = 100%

Feedback:

Remember, the two things needed to get a p-value are a test statistic, in this case a chi-squared statistic equal to **4.9575**, and a distribution of the test statistic, in this case a chi-squared distribution with 2 degrees of freedom, i.e., df=2. Going to the "Making Inference" page of the textbook and looking up the chi-squared distribution with 2 degrees of freedom, we see the value of 5 isn't too likely, but has some height to the curve above that value. Looking out to the right, we might guess around 5% or so for the p-value, and since 0.0838 is the only value close to that, we guess that as our answer.

Question14: *Score 1/1*

Use the KidsFeet dataset in R to calculate the predicted probability that the birthyear of a child is '88 when the length of the foot (measured in October of 1997) is 26 cm long. (Scroll down to select your answer.)



Your response	Correct response
0.709	

Auto graded Grade: 1/1.0

Total grade: 1.0×1/1 = 100%

Feedback:

```
> kids.glm <- glm(birthyear==88 ~ length, data=KidsFeet, family=binomial)

> predict(kids.glm, data.frame(length=26), type="response")
```

Question15: *Score 1/1*

Use the **KidsFeet** dataset in R to perform a test to determine if β_1 is significantly different from zero in the model shown below.

Report the p-value of your test.

$$\underbrace{Y_i}_{\text{width}} = \beta_0 + \beta_1 \underbrace{X_i}_{\text{birthmonth}} + \epsilon_i$$

Your response	Correct response
P-value = 0.995	

Auto graded Grade: 1/1.0

Total grade: 1.0×1/1 = 100%

Feedback:

```
plot(width ~ birthmonth, data=KidsFeet)
kids.lm <- lm(width ~ birthmonth, data=KidsFeet)
summary(kids.lm)

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  8.9932912  0.1730841  51.959   <2e-16 ***
birthmonth  -0.0001612  0.0249161  -0.006    0.995
```

Question16: *Score 1/1*

Select the best interpretation of the slope estimate from the logistic regression summary(...) output shown below.

```
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) 1.364569   1.126934   1.211   0.2259
X          -0.014717   0.007121  -2.067   0.0388 *
```

Your response	Correct response
Every 1 unit increase in X results in a 1.46% drop in the odds that $Y_i = 1$.	

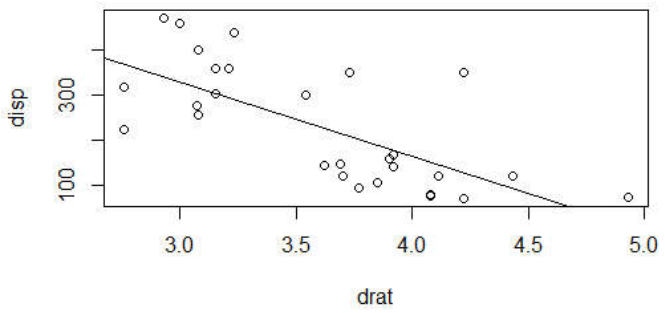
Auto graded Grade: 1/1.0

Total grade: 1.0×1/1 = 100%
Feedback:

Using e^b1, exp(-0.014717) = 0.9853908, shows that the odds are 98.54% of what they were for every 1 unit increase in X. This means the odds are decreasing by 1-0.9853908 = 0.01460923, or 1.46%.

Question17: *Score 1/1*

Which of the following graphics correctly depicts the results of an analysis that uses the **disp** and **drat** columns of the **mtcars** dataset?

Your response	Correct response
	

Auto graded **Grade:** 1/1.0

Total grade: 1.0×1/1 = 100%
Feedback:

First, notice that disp and drat are both quantitative. So the analysis that would make the most sense is a linear regression, which results in the scatterplot and regression line shown. The code for this is given by

```
> mt.lm <- lm(disp ~ drat, data=mtcars)
> plot(disp ~ drat, data=mtcars)
> abline(mt.lm)
```

None of the other graphics can be created directly from the data.

Question18: *Score 1/1*

A student tried to perform the following t Test in R, but got an error message. Go ahead and run the code yourself to see the error message.

```
t.test(mpg ~ cyl, data=mtcars)
```

Which of the following would provide the student with a solution to this error message?

Your response	Correct response
Try using kruskal.test(...) instead of t.test(...).	

Auto graded **Grade:** 1/1.0

Total grade: 1.0×1/1 = 100%
Feedback:

Try each of the solutions and you will see that only kruskal.test(...) fixes the problem. The reason for this is that cyl has 4, 6, and 8 as values, which is creating three groups in the data. A t.test(...) and wilcox.test(...) are only for two groups. So moving to a kruskal.test(...) allows all three groups to be included in the test.

Question19: *Score 1/1*

Run the following codes in R.

View(mtcars)

?mtcars

If a researcher were to use only the **hp** and **vs** columns, which type of analysis would NOT make sense to try and perform with this data as it is currently shown?

Your response	Correct response
A chi-squared Test.	

Auto graded Grade: 1/1.0

Total grade: 1.0×1/1 = 100%
Feedback:

A logistic regression could be used to predict vs==1 based on the hp value.

A Wilcoxon Rank Sum Test could be used to compare the median hp for the vs==1 and vs==0 groups.

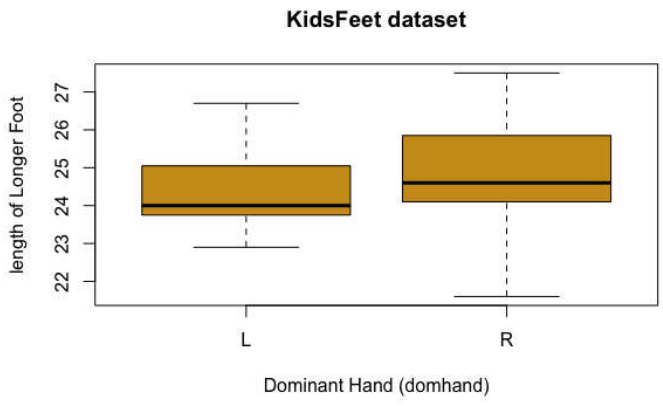
An Independent Samples t Test could be used to compare the mean hp for the vs==1 and vs==0 groups.

A chi-squared test would not work with this data as currently shown because hp is a quantitative variable. Of course, hp could be cut into interesting groups using say the cut(...) function, and then a chi-squared test performed. But with the data as it currently stands, a chi-squared test would not make sense.

Question20: *Score 1/1*

Make a boxplot in R which depicts the length of feet for right-handed and left-handed kids as recorded in the KidsFeet dataset.

Select the graphic you obtain from the options below.

Your response	Correct response
	

Auto graded Grade: 1/1.0

Total grade: 1.0×1/1 = 100%

Feedback:

boxplot(length ~ domhand, data=KidsFeet, main="KidsFeet dataset", ylab="length of Longer Foot", xlab="Dominant Hand (domhand)", col="goldenrod3")

Question21: *Score 1/1*

A certain hospital collected data from mothers who recently gave birth. They asked the mother if she smoked at all during the pregnancy and recorded her answer as either "yes" or "no". They also recorded the birth weight of the newborn child in grams.

If a logistic regression were used with this data, which of the following would properly word the question for the analysis?

Your response	Correct response
Can the birth weight of the child predict whether or not the mother smoked during the pregnancy?	

Auto graded Grade: 1/1.0

Total grade: 1.0×1/1 = 100%

Feedback:

If an independent samples t test (or Wilcoxon Rank Sum Test) were used for this data, then the question would be either of

Does smoking during a pregnancy effect the birth weight of the child?

Are children born to mothers that smoke lighter on average than children born to mothers who don't smoke?

However, since it is a logistic regression, then we are using a quantitative x-variable (birth weight in this case) to predict a binomial outcome (whether or not the mother smoked in this case). Thus, the correct answer would be

Can the birth weight of the child predict whether or not the mother smoked during the pregnancy?

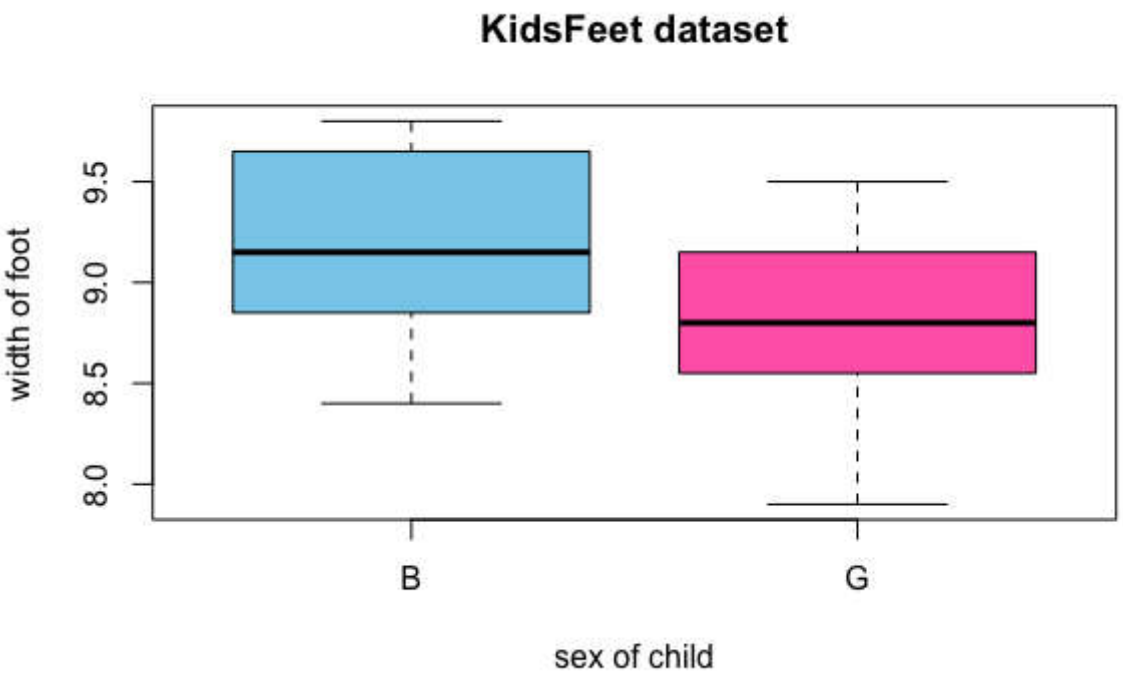
This question is just jibberish:

Does the probability of the child's birth weight depend on whether or not the mother smoked during the pregnancy?

What does "probaiblity of the child's birth weight" even mean? It would need to say something like "probability of the child weighing more than 3,000 grams" to be meaningful.

Question22: *Score 1/1*

Perform an appropriate hypothesis test in R to decide if the medians displayed in the graph are significantly different or not.



Your response	Correct response
Yes, the median foot width of boys of 9.15 cm **is** significantly longer than the median width for girls of 8.80 cm (p-value = 0.02004).	

Auto graded **Grade:** 1/1.0

Total grade: 1.0×1/1 = 100%

Feedback:

Because the answers all speak about medians, and we are comparing two groups, the appropriate test is the Wiclxon Rank Sum Test:

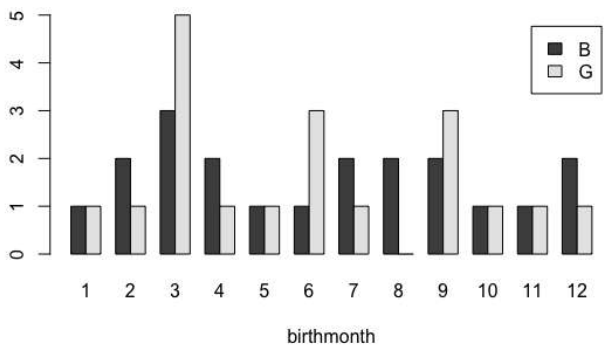
```
wilcox.test(width ~ sex, data=KidsFeet)

    Wilcoxon rank sum test with continuity correction

data:  width by sex
W = 273, p-value = 0.02004
alternative hypothesis: true location shift is not equal to 0
```

Question23: *Score 1/1*

Use the KidsFeet dataset in R. Which of the following correctly displays how many boys (B) and girls (G) from the dataset were born in each month of the year? (Jan = 1, ..., Dec = 12)

Your response	Correct response
	

Auto graded Grade: 1/1.0

Total grade: 1.0×1/1 = 100%

Feedback:

barplot(table(KidsFeet$sex, KidsFeet$birthmonth), beside=TRUE, legend=TRUE, xlab="birthmonth")

Or just

table(KidsFeet$sex, KidsFeet$birthmonth)

Question24: *Score 1/1*

What two things are needed to compute a p-value?

Your response	Correct response
A test statistic and a distribution of the test statistic.	

Auto graded Grade: 1/1.0

Total grade: 1.0×1/1 = 100%

Question25: *Score 1/1*

Use the KidsFeet dataset in R to calculate the average foot **length** of the **boys** in the dataset who were born in **January**.

Your response	Correct response
25.2	

Auto graded Grade: 1/1.0

Total grade: 1.0×1/1 = 100%

Feedback:

Many ways you could do this. Here are a couple quick ways:

```
mean(length ~ sex, data=subset(KidsFeet, birthmonth==1))
mean(length ~ sex, data=KidsFeet, groups=birthmonth)
```