STA130 Rstudio Homework

Problem Set 6

[Student Name] ([Student Number]), with Josh Speagle & Scott Schwartz

Instructions

Complete the exercises in this `.Rmd` file and submit your `.Rmd` and knitted `.pdf` output through Quercus by 11:59 pm E.T. on Thursday, March 9.

```
library(tidyverse)
```

Question 1: Broadway, the Musical

Lin-Manuel Miranda was nominated for "Best Original Song" for the March 27, 2022 the Academy Awards (also known as the Oscars) for his work on the Disney movie Encanto. Miranda had already won an Emmy, Grammy, and Tony (mostly for his work on the broadway musical "Hamilton"), so he was very close to the (EGOT)[https://www.vanityfair.com/hollywood/2022/02/oscar-nominations-2022-will-lin-manuel-miranda-finally-egot-for-encanto] (Emmy, Grammy, Oscar and Tony), a rare occurrence as only 16 people have won all four awards see here.

Unfortunately, Miranda did not win the Oscar in 2022. Perhaps he will soon!

In this question, we will look at a sample of weekly Broadway musical data available in the `broadway.csv`. This data set contains a sample of Broadway musical information for 500 weeks from 1985 to 2020. In this data set, an observation is one Broadway musical in a particular week (ending on a Sunday). Variables of interest are:

- `show`: Name of the Broadway musical/show.
- `Hamilton`: indicates whether the musical is Hamilton or not.
- `week_ending`: Date of the end of the weekly measurement period. Always a Sunday.
- `weekly_gross_overall`: Weekly box office gross for all shows.
- `avg_ticket_price`: Average price of tickets sold in a particular week.
- `top_ticket_price`: Highest price of tickets sold in a particular week.
- `seats_sold`: Total seats sold for all performances and previews in a particular week.
- `pct_capacity`: Percent of theater capacity sold. Shows can exceed 100% capacity by selling standing room tickets.

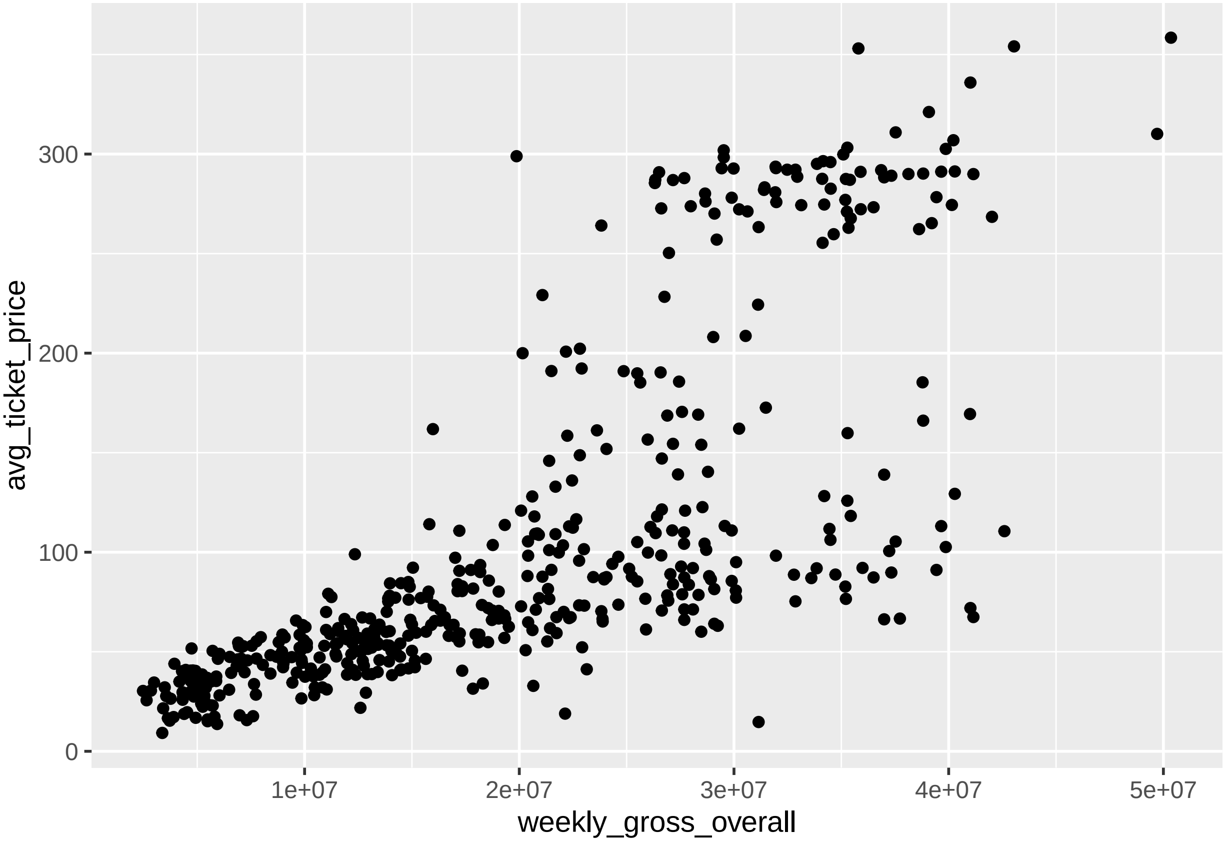
```
# load in data
broadway_data <- read_csv("broadway.csv")

# preview data
glimpse(broadway_data)
```

```
## Rows: 500
## Columns: 8
## $ show                    <chr> "La Cage aux Folles", "42nd Street", "42nd Street~
## $ Hamilton                <chr> "No", "No", "No", "No", "No", "No", "No", "No", "~
## $ week_ending             <date> 1985-07-28, 1985-09-08, 1985-09-15, 1985-12-15, ~
## $ weekly_gross_overall <dbl> 2989271, 2474396, 2844860, 4169643, 3555363, 3632~
## $ avg_ticket_price     <dbl> 34.54, 30.31, 30.50, 35.00, 27.74, 16.60, 17.19, ~
## $ top_ticket_price     <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, N~
## $ seats_sold           <dbl> 11841, 7251, 7890, 10846, 2803, 2204, 5740, 10861~
## $ pct_capacity         <dbl> 0.8795, 0.5477, 0.5959, 0.8056, 0.2967, 0.4364, 0~
```

In this question, we will explore different ways to estimate the average ticket price for Broadway shows.

(a) Make a **scatter plot** showing the relationship between the average ticket price (on the y-axis) and the weekly gross overall sales (on the x-axis).

```
ggplot(broadway_data) + aes(x = weekly_gross_overall, y = avg_ticket_price) + geom_point()
```



In 1-2 sentences, explain whether or not you think it is appropriate to characterize and summarize the association in the above plot with a straight line.

> *A line wouldn't be good to summarize the association as the the residuals to the line would be big for the end of the graph.*

(b) Use the `mutate()` function to add the new variables `log_avg_ticket_price = log10(avg_ticket_price)` and `weekly_gross_overall_mil=weekly_gross_overall/1e6` to the data set.

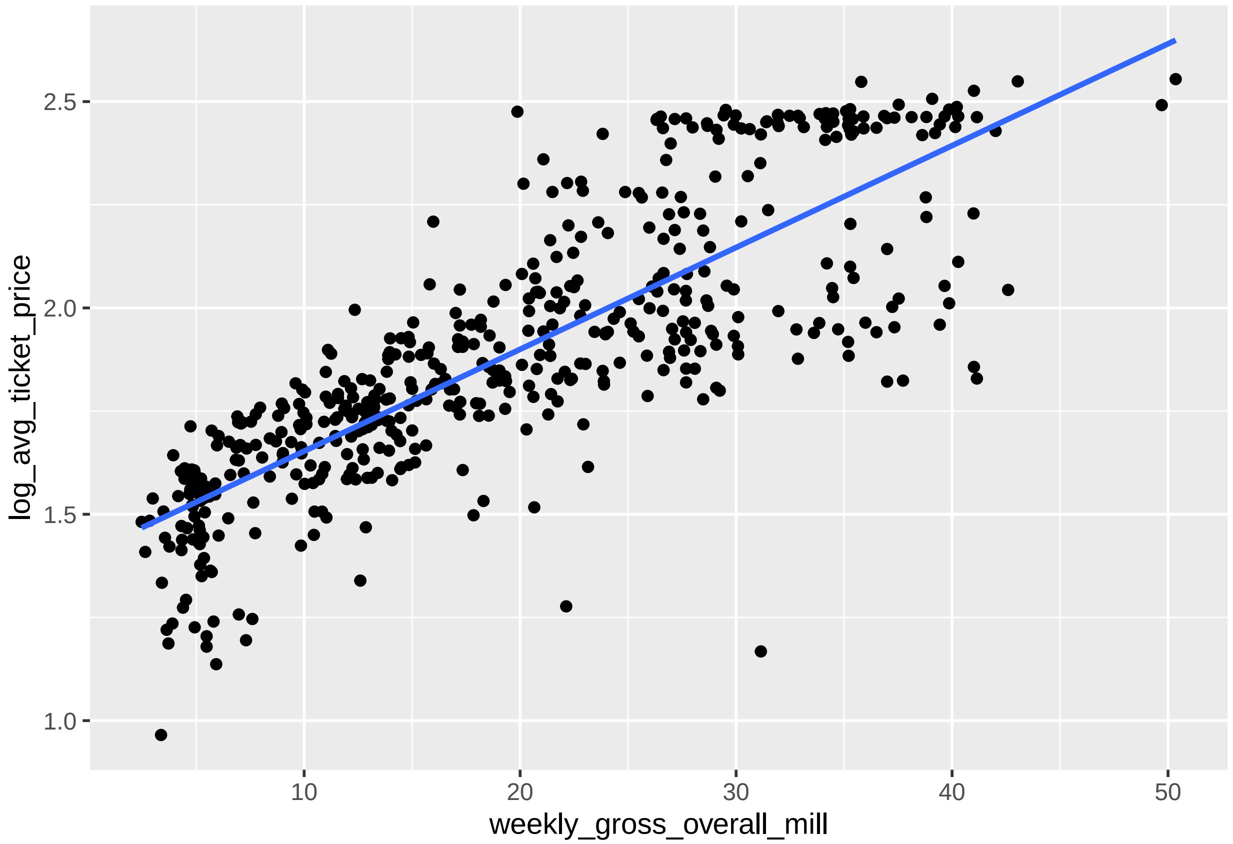
*Note: Based on the dataset(s) you are working with on the capstone project, you may already be experimenting with **transforming variables** to improve the behaviour of your modelling approach and/or quality of your predictions. You will likely learn more about transforming variables in future courses.*

```
new_broadway_data <-broadway_data %>%
  mutate(log_avg_ticket_price = log10(avg_ticket_price), weekly_gross_overall_mill = weekly_gross_overal
new_broadway_data
```

```
## # A tibble: 500 x 10
##    show       Hamil~1 week_end~2 weekl~3 avg_t~4 top_t~5 seats~6 pct_c~7 log_a~8
##    <chr>      <chr>   <date>       <dbl>   <dbl>   <dbl>   <dbl>   <dbl>   <dbl>
##  1 La Cage a~ No      1985-07-28 2989271    34.5      NA   11841   0.880    1.54
##  2 42nd Stre~ No      1985-09-08 2474396    30.3      NA    7251   0.548    1.48
##  3 42nd Stre~ No      1985-09-15 2844860    30.5      NA    7890   0.596    1.48
##  4 La Cage a~ No      1985-12-15 4169643    35        NA   10846   0.806    1.54
##  5 The Odd C~ No      1986-01-26 3555363    27.7      NA    2803   0.297    1.44
##  6 Loot       No      1986-04-06 3632735    16.6      NA    2204   0.436    1.22
##  7 Arsenic a~ No      1986-06-29 3900725    17.2      NA    5740   0.584    1.24
##  8 The Myste~ No      1986-07-13 3486170    32.1      NA   10861   0.953    1.51
##  9 Arsenic a~ No      1986-07-20 3716807    15.4      NA    9592   0.854    1.19
## 10 I'm Not R~ No      1986-09-14 3762479    26.4      NA    6011   0.960    1.42
## # ... with 490 more rows, 1 more variable: weekly_gross_overall_mill <dbl>, and
## #   abbreviated variable names 1: Hamilton, 2: week_ending,
## #   3: weekly_gross_overall, 4: avg_ticket_price, 5: top_ticket_price,
## #   6: seats_sold, 7: pct_capacity, 8: log_avg_ticket_price
```

Now plot the association between `log_avg_ticket_price` (on the y-axis) and `weekly_gross_overall_mil` (on the x-axis) and use `geom_smooth(method=lm, se=FALSE)` to add a **line of best fit** to the plot.

```
ggplot(new_broadway_data) + aes(x = weekly_gross_overall_mill, y = log_avg_ticket_price) + geom_point()
```



In 2-4 sentences, describe the association you observe in the plot and whether the transformation to `log_avg_ticket_price` and/or `weekly_gross_overall_mil` was helpful or not.

> *It was very helpful as the points are more correlated. The total residual will probably also be less than the previous one.*

(c) Use the `cor()` function to calculate the **correlation** between `log_avg_ticket_price` and `weekly_gross_overall_100k`.

Hint: Remember that you can access individual variables/columns in a tibble using the syntax `tibble$variable`*.*

```
cor(new_broadway_data$log_avg_ticket_price, new_broadway_data$weekly_gross_overall)
```

```
## [1] 0.8154224
```

In 1-2 sentences, discuss whether this number implies `log_avg_ticket_price` and `weekly_gross_overall_mil` are strongly/weakly/not at all positively/negatively correlated.

> *Considering that 1 is a high correlation and 0 is no correlation, a 0.8 correlation is pretty good.*

(d) Write down a simple **linear regression model** with a **response variable** y corresponding to `log_avg_ticket_price` and an **explanatory variable** x corresponding to `weekly_gross_overall_mil`.

Hint: A reminder that if you math equations or other symbols directly from another source into your .Rmd document, you may get errors when trying to knit. Instead, try and use `$` *notation to write equations. A single* `$y=a$` *will get you math within text, while* `$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1$$` *will put your equation on a new line by itself. A few useful symbols here may include epsilon (ϵ), "not equal" ($\neq$), superscripts (e.g. i^{th}), and subscripts (e.g. i_{th}).*

> $y = log_a vg_t icket_p ricex$*=weekly_gross_overall_mil%)*

Now explain each component of the model above.

> *y is the dependent variable(response variable), x is the independent variable(explanatory variable), B1 would be how much the response variable changes based on the explanatory variable and B0 is the y intercept of the slope.*

(e) State the **null and alternative hypotheses** you would use to assess whether the slope of the linear regression model where `weekly_gross_overall_100k` is predicting `log_avg_ticket_price`.

> *null: weekly_gross_overall_100k doesn't have an impact on log_avg_ticket_price alternative hypothesis: weekly_gross_overall_100k has an impact on log_avg_ticket_price*

(f) Use the `lm()` function to find the line of best fit for your simple linear regression model and provide a summary of the results by piping your output into the `summary()` function.

Hint: Please remember to check on the format of the input arguments for `lm()`*, since they are different from most of the functions we are have previously dealt with.*

In 3-6 sentences, interpret the different rows/columns/entries from the `summary()` output in the context of the underlying data and model.

Hint: In addition to information on the course slides, you may find this post helpful to interpret all the different parts of the summary output.

> *REPLACE THIS TEXT WITH YOUR ANSWER*

Using an α significance level of $\alpha = 10^{-3}$, draw a conclusion regarding the hypothesis test you defined earlier related to the inferred slope.

> *REPLACE THIS TEXT WITH YOUR ANSWER*

Question 2: Hamilton

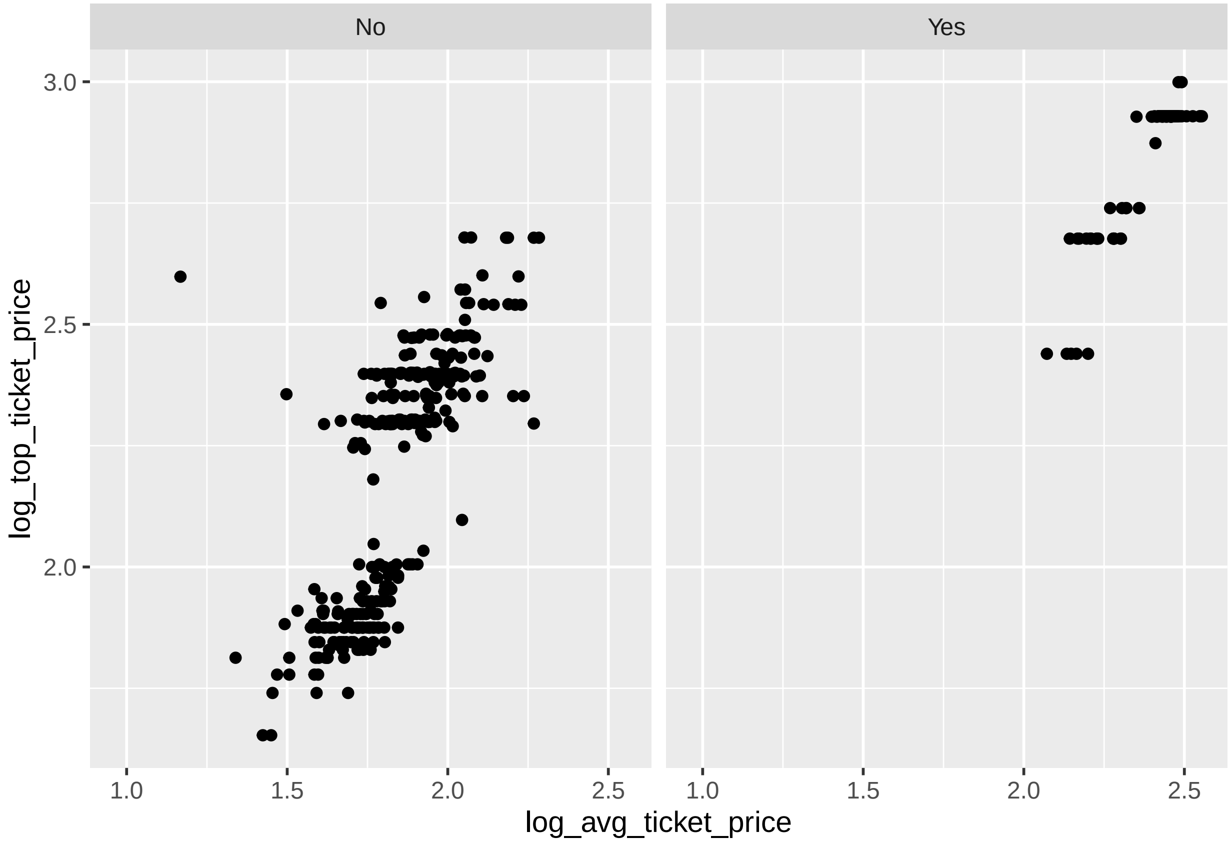
(a) Use `mutate()` to create a new column, `log_top_ticket_price`, the same way you created `log_avg_ticket_price`. Then, make a scatter plot of the association between `log_top_ticket_price` (on the y-axis) and `log_avg_ticket_price` (on the x-axis) **faceted** by whether the musical was "Hamilton" or not.

Hint: Using `ggplot`, adding `+ facet_wrap(~ Hamilton)` to the options is an easy way to facet the data.

```
newest <-new_broadway_data %>%
  mutate(log_top_ticket_price = log10(top_ticket_price))
newest
```

```
## # A tibble: 500 x 11
##    show       Hamil~1 week_end~2 weekl~3 avg_t~4 top_t~5 seats~6 pct_c~7 log_a~8
##    <chr>      <chr>   <date>       <dbl>   <dbl>   <dbl>   <dbl>   <dbl>   <dbl>
##  1 La Cage a~ No      1985-07-28 2989271    34.5      NA   11841   0.880    1.54
##  2 42nd Stre~ No      1985-09-08 2474396    30.3      NA    7251   0.548    1.48
##  3 42nd Stre~ No      1985-09-15 2844860    30.5      NA    7890   0.596    1.48
##  4 La Cage a~ No      1985-12-15 4169643    35        NA   10846   0.806    1.54
##  5 The Odd C~ No      1986-01-26 3555363    27.7      NA    2803   0.297    1.44
##  6 Loot       No      1986-04-06 3632735    16.6      NA    2204   0.436    1.22
##  7 Arsenic a~ No      1986-06-29 3900725    17.2      NA    5740   0.584    1.24
##  8 The Myste~ No      1986-07-13 3486170    32.1      NA   10861   0.953    1.51
##  9 Arsenic a~ No      1986-07-20 3716807    15.4      NA    9592   0.854    1.19
## 10 I'm Not R~ No      1986-09-14 3762479    26.4      NA    6011   0.960    1.42
## # ... with 490 more rows, 2 more variables: weekly_gross_overall_mill <dbl>,
## #   log_top_ticket_price <dbl>, and abbreviated variable names 1: Hamilton,
## #   2: week_ending, 3: weekly_gross_overall, 4: avg_ticket_price,
## #   5: top_ticket_price, 6: seats_sold, 7: pct_capacity,
## #   8: log_avg_ticket_price
```

```
ggplot(newest) + aes(x= log_avg_ticket_price, y = log_top_ticket_price) + facet_wrap(~Hamilton) + geom_
```



(b) Calculate the correlation between `log_top_ticket_price` and `log_avg_ticket_price` for both Hamilton and non-Hamilton musicals.

Hint: You might find `group_by()` and `summarize()` to be helpful here. Also, remember to be on the lookout for `NA` values.

```
ham <-filter(newest, Hamilton == "Yes")
ham
```

```
## # A tibble: 100 x 11
##    show     Hamilton week_ending weekl~1 avg_t~2 top_t~3 seats~4 pct_c~5 log_a~6
##    <chr>    <chr>    <date>        <dbl>   <dbl>   <dbl>   <dbl>   <dbl>   <dbl>
##  1 Hamilton Yes      2015-08-02   2.88e7    140.     275   10619    1.00    2.15
##  2 Hamilton Yes      2015-08-09   2.64e7    118.     275   10638    1.01    2.07
##  3 Hamilton Yes      2015-08-23   2.25e7    136.     275   10708    1.01    2.13
##  4 Hamilton Yes      2015-09-06   2.22e7    159.     275   10706    1.01    2.20
##  5 Hamilton Yes      2015-09-13   2.14e7    146.     275   10703    1.01    2.16
##  6 Hamilton Yes      2015-10-11   2.60e7    157.     475   10717    1.01    2.19
##  7 Hamilton Yes      2015-10-25   2.74e7    139.     475   10708    1.01    2.14
##  8 Hamilton Yes      2015-11-01   2.28e7    149.     475   10726    1.01    2.17
##  9 Hamilton Yes      2015-11-08   2.66e7    147.     475   12050    1.01    2.17
## 10 Hamilton Yes      2016-01-24   1.60e7    162.     475    8062    1.02    2.21
## # ... with 90 more rows, 2 more variables: weekly_gross_overall_mill <dbl>,
## #   log_top_ticket_price <dbl>, and abbreviated variable names
## #   1: weekly_gross_overall, 2: avg_ticket_price, 3: top_ticket_price,
## #   4: seats_sold, 5: pct_capacity, 6: log_avg_ticket_price
```

```
not_ham <- newest %>% filter(Hamilton == "No")
not_ham<-na.omit(not_ham)
not_ham
```

```
## # A tibble: 306 x 11
##    show       Hamil~1 week_end~2 weekl~3 avg_t~4 top_t~5 seats~6 pct_c~7 log_a~8
##    <chr>      <chr>   <date>       <dbl>   <dbl>   <dbl>   <dbl>   <dbl>   <dbl>
##  1 Kiss of t~ No      1995-06-11  9.01e6    44.1    70      4372   0.483    1.64
##  2 Beauty an~ No      1995-08-06  7.54e6    53.0    67.5   13282   0.954    1.72
##  3 The Phant~ No      1995-09-03  6.91e6    54.6    67.5   13080   1.02     1.74
##  4 Les Miser~ No      1995-09-17  6.97e6    42.7    67.5   10031   0.887    1.63
##  5 How to Su~ No      1995-12-31  1.29e7    52.4    67.5   10629   0.973    1.72
##  6 Show Boat  No      1996-01-07  7.96e6    57.3    75     10991   0.716    1.76
##  7 A Midsumm~ No      1996-04-07  9.00e6    42.2    65      6813   0.588    1.63
##  8 A Midsumm~ No      1996-05-19  1.08e7    39.6    65      6424   0.555    1.60
##  9 Les Miser~ No      1996-06-09  9.86e6    46.0    70     10632   0.941    1.66
## 10 A Funny T~ No      1996-07-28  8.97e6    58.6    70     10562   0.816    1.77
## # ... with 296 more rows, 2 more variables: weekly_gross_overall_mill <dbl>,
## #   log_top_ticket_price <dbl>, and abbreviated variable names 1: Hamilton,
## #   2: week_ending, 3: weekly_gross_overall, 4: avg_ticket_price,
## #   5: top_ticket_price, 6: seats_sold, 7: pct_capacity,
## #   8: log_avg_ticket_price
```

```
na.omit(ham)
```

```
## # A tibble: 100 x 11
##    show     Hamilton week_ending weekl~1 avg_t~2 top_t~3 seats~4 pct_c~5 log_a~6
##    <chr>    <chr>    <date>        <dbl>   <dbl>   <dbl>   <dbl>   <dbl>   <dbl>
##  1 Hamilton Yes      2015-08-02   2.88e7    140.     275   10619    1.00    2.15
##  2 Hamilton Yes      2015-08-09   2.64e7    118.     275   10638    1.01    2.07
##  3 Hamilton Yes      2015-08-23   2.25e7    136.     275   10708    1.01    2.13
##  4 Hamilton Yes      2015-09-06   2.22e7    159.     275   10706    1.01    2.20
##  5 Hamilton Yes      2015-09-13   2.14e7    146.     275   10703    1.01    2.16
##  6 Hamilton Yes      2015-10-11   2.60e7    157.     475   10717    1.01    2.19
##  7 Hamilton Yes      2015-10-25   2.74e7    139.     475   10708    1.01    2.14
##  8 Hamilton Yes      2015-11-01   2.28e7    149.     475   10726    1.01    2.17
##  9 Hamilton Yes      2015-11-08   2.66e7    147.     475   12050    1.01    2.17
## 10 Hamilton Yes      2016-01-24   1.60e7    162.     475    8062    1.02    2.21
## # ... with 90 more rows, 2 more variables: weekly_gross_overall_mill <dbl>,
## #   log_top_ticket_price <dbl>, and abbreviated variable names
## #   1: weekly_gross_overall, 2: avg_ticket_price, 3: top_ticket_price,
## #   4: seats_sold, 5: pct_capacity, 6: log_avg_ticket_price
```

```
cor(not_ham$log_top_ticket_price, not_ham$log_avg_ticket_price)
```

```
## [1] 0.757476
```

```
cor(ham$log_avg_ticket_price, ham$log_top_ticket_price)
```

```
## [1] 0.9292493
```

Write 1-2 sentences discussing what the correlations you computed above imply in terms of how much `log_top_ticket_price` and `log_avg_ticket_price` relate to each other and whether there are any big differences between whether the musical was Hamilton or not.

> *When the musical was hamilton the average ticket price and top ticket price were extremely correlated which logically makes sense as they're to the same show. For non-hamilton shows*

> *however, its slightly less which also makes sense as there are many other shows and to assume their average grows alongside their top ticket price for all the shows would be irrational.*

(c) Find the lines of best fit for a simple linear regression model for the Hamilton and non-Hamilton musicals, respectively. Then provide a summary of the results by piping your output(s) into the `summary()` function.

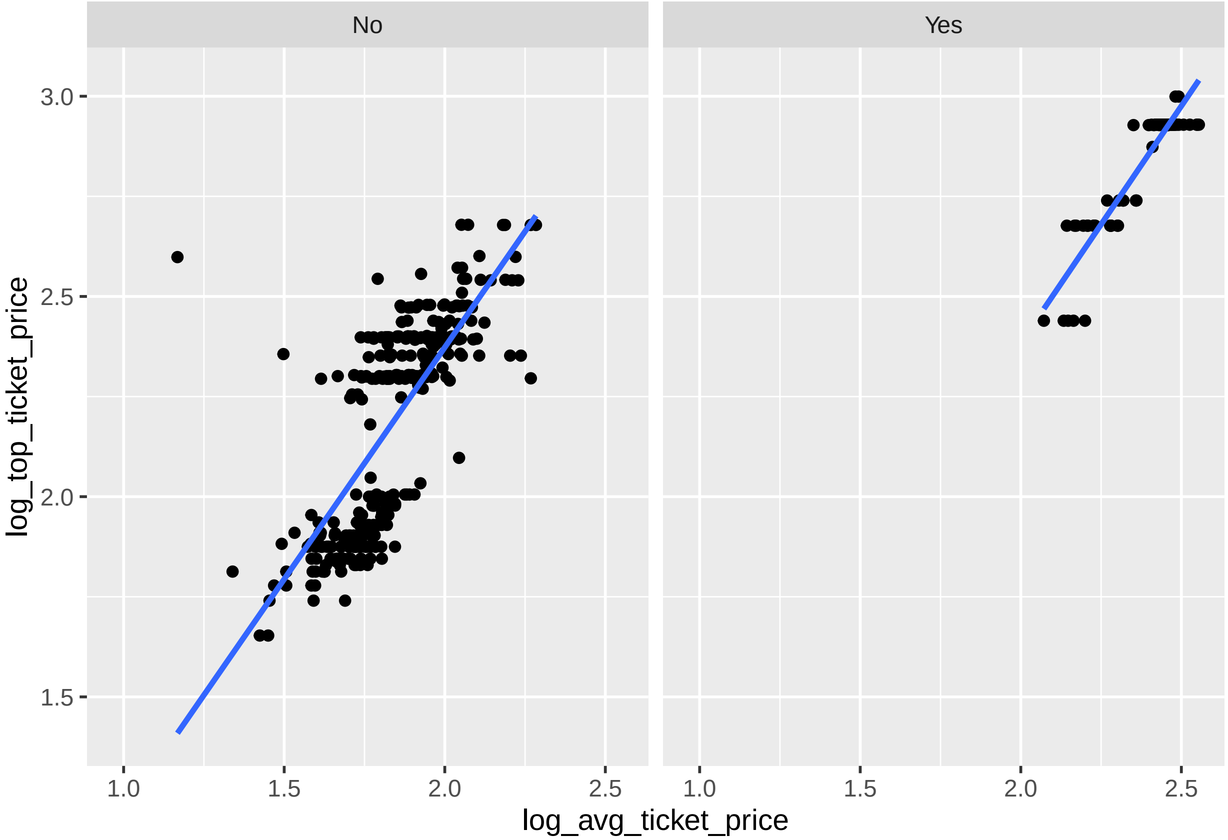
```
#lm(ham ~ not_ham, newest)
```

In 2-3 sentences, please comment on what the fitted coefficients (slope and intercept) of your model implies for the relationship between `log_top_ticket_price` and `log_avg_ticket_price`. Based on the estimated standard errors, do you think the fitted coefficients of the two models are meaningfully different?

> *REPLACE THIS TEXT WITH YOUR ANSWER*

(d) Plot the association between `log_top_ticket_price` (on the y-axis) and `log_avg_ticket_price` (on the x-axis) split up by `Hamilton` using `facet_wrap()` and with the line of best fit added to both panels using `geom_smooth(method=lm, se=FALSE)`.

```
ggplot(newest) + aes(x = log_avg_ticket_price, y =log_top_ticket_price) + facet_wrap(~Hamilton) + geom_p
```



Question 3: Starbucks

The `starbucks.csv` dataset contains data on calories and carbohydrates (in grams) in Starbucks food menu items.

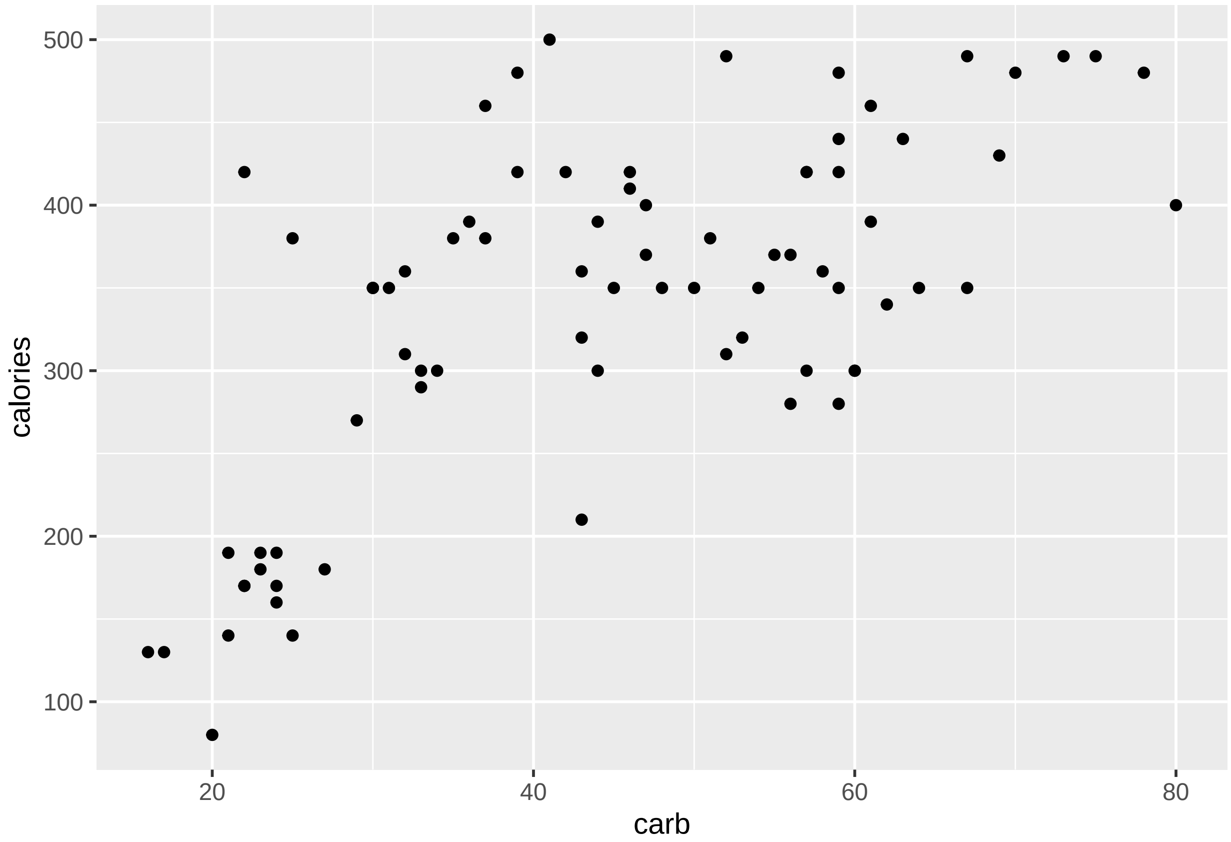
```
# load in data
starbucks_data <- read_csv("starbucks.csv")
```

```
# preview data
glimpse(starbucks_data)
```

```
## Rows: 77
## Columns: 7
## $ item     <chr> "8-Grain Roll", "Apple Bran Muffin", "Apple Fritter", "Banana~
## $ calories <dbl> 350, 350, 420, 490, 130, 370, 460, 370, 310, 420, 380, 320, 3~
## $ fat      <dbl> 8, 9, 20, 19, 6, 14, 22, 14, 18, 25, 17, 12, 17, 21, 5, 18, 1~
## $ carb     <dbl> 67, 64, 59, 75, 17, 47, 61, 55, 32, 39, 51, 53, 34, 57, 52, 7~
## $ fiber    <dbl> 5, 7, 0, 4, 0, 5, 2, 0, 0, 0, 2, 3, 2, 2, 3, 3, 2, 3, 0, 2, 0~
## $ protein  <dbl> 10, 6, 5, 7, 0, 6, 7, 6, 5, 7, 4, 6, 5, 5, 12, 7, 8, 6, 0, 10~
## $ type     <chr> "bakery", "bakery", "bakery", "bakery", "bakery", "bakery", "~
```

(a) Produce a plot that shows the association between carbohydrates (y-axis) and calories (x-axis) in Starbucks menu items.

```
starbucks_data %>%
  ggplot() + aes(x=carb, y=calories) +
  geom_point()
```



Write 1-2 sentences describing any association you observe.

Carbs and Calories are not extremely correlated

(b) Estimate the correlation coefficient between carbohydrates and calorie content in Starbucks menu items based on the plot you produced above *entirely by eye* (i.e. without actually computing anything). Write and then justify your answer below.

0.3

Now calculate the correlation between carbohydrate and calorie content of Starbucks menu items.

```
cor(starbucks_data$carb, starbucks_data$calories)
```

```
## [1] 0.674999
```

How does this compare to your earlier "by eye" estimate?

It makes sense as they're not extremely correlated but logically increasing carbs also means increasing calories. There are other factors that increase calories in a Starbucks drink as well.

(c) Fit a simple linear regression model where `calories` is the response variable and `carb` is the explanatory variable to these data. Describe the main results highlighted in the `summary()` output in 2-3 sentences.

```
line <-lm(calories ~ carb, starbucks_data)
summary(line)
```

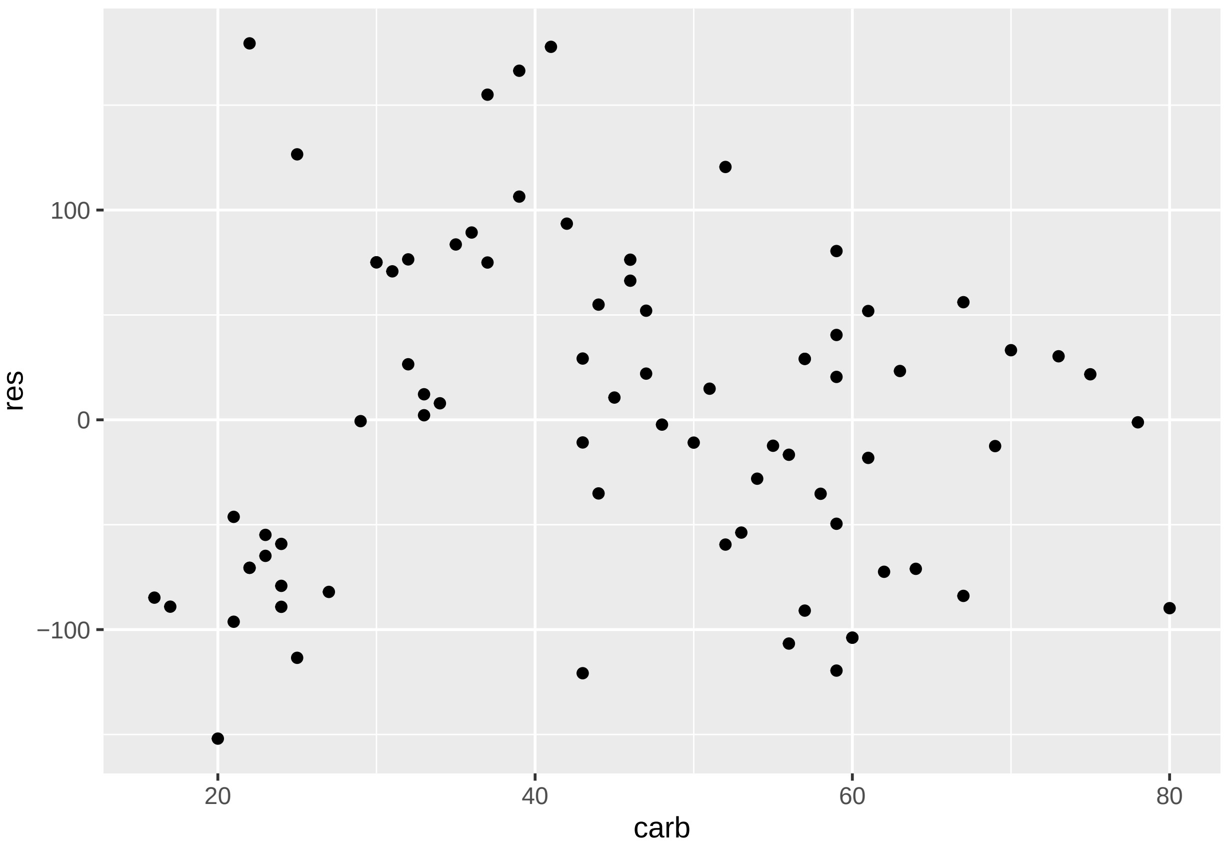
```
##
## Call:
## lm(formula = calories ~ carb, data = starbucks_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -151.962  -70.556   -0.636   54.908  179.444
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 146.0204    25.9186   5.634 2.93e-07 ***
## carb          4.2971     0.5424   7.923 1.67e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 78.26 on 75 degrees of freedom
## Multiple R-squared:  0.4556, Adjusted R-squared:  0.4484
## F-statistic: 62.77 on 1 and 75 DF,  p-value: 1.673e-11
```

There are 77 residuals, 8 coefficients, 3 data frames, and 1 r-squared value. The 77 residuals shed light on the deviations from the line of best fit, which would be 77 deviations.

(d) Based on the estimated line of best fit computed above, calculate/extract the fitted residuals $\epsilon_1, \ldots, \epsilon_n$ and plot them as a function of the explanatory variable `carb`.

Hint: The output of the `lm()` function might be handy here. Try `?lm` to get some additional information on the values that are returned.

```
res <-residuals(line)
line %>% ggplot(aes(x=carb, y= res)) + geom_point()
```



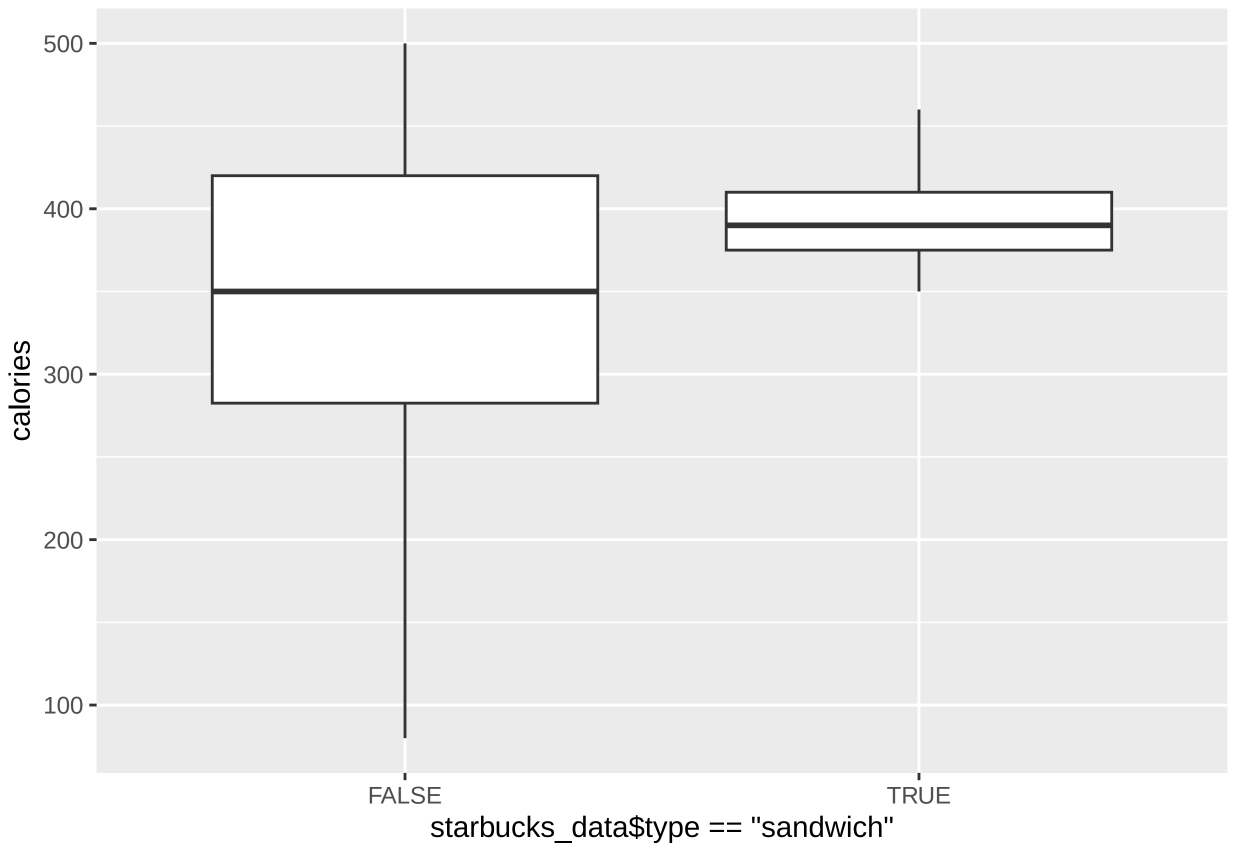
In 1-2 sentences, comment on any trends (or lack of trends) that you may observe and what this implies about the overall fitted relationship.

REPLACE THIS TEXT WITH YOUR ANSWER

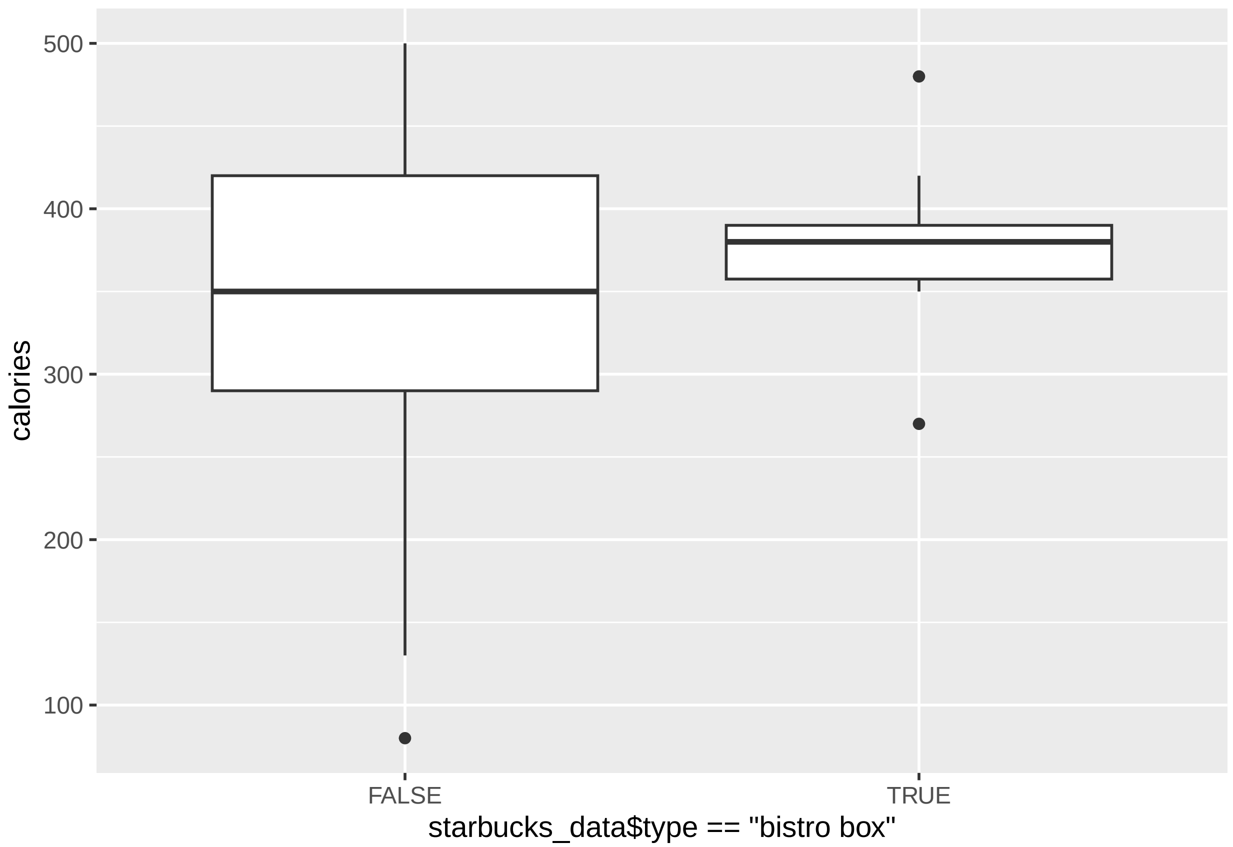
Question 4: No Free Lunch

(a) Based on the Starbucks data, create a new data set called `starbucks_lunch` which only contains food items of the `"sandwich"` or `"bistro box"` in `type`. Then create a box plot comparing the distribution of calories for these two types of items along with a summary table containing the total number of objects in each group along with their respective mean calories.

```
starbucks_lunch <- filter(starbucks_data, type == "sandwich" | type =="bistro box")
ggplot(starbucks_data) + aes(x = starbucks_data$type == "sandwich", y =calories) + geom_boxplot()
```



```
ggplot(starbucks_data) + aes(x = starbucks_data$type == "bistro box", y =calories) + geom_boxplot()
```



```
sandwich <- filter(starbucks_lunch, type == "sandwich")
bistro <- filter(starbucks_lunch, type == "bistro box")
summarize(sandwich,
          mean(calories),
          nrow(sandwich))
```

```
## # A tibble: 1 x 2
##   `mean(calories)` `nrow(sandwich)`
##              <dbl>            <int>
## 1             396.                7
```

```
summarize(bistro,
          mean(calories),
          nrow(bistro))
```

```
## # A tibble: 1 x 2
##   `mean(calories)` `nrow(bistro)`
##              <dbl>          <int>
## 1             378.              8
```

(b) Write down a simple **linear regression model** with a **response variable** y corresponding to `calories` and an **explanatory variable** x corresponding to an binary **indicator variable** as a function of `type`. In other words, x takes values of 1 or 0 and is defined as:

$$x = \begin{cases} 1 \text{ if `type` = `sandwich`} \\ 0 \text{ if `type` = `bistrobox`} \end{cases}$$

Note that this is equivalent to coercing `type == "sandwich"` to an integer value.

> *1: y(mean calories of sandwich) = x(1) 0: y(mean calories of bistrobox = x(0)*

Now explain each component of the model above. Note that your interpretation should involve the mean calories for items in each respective group.

> *y is calories, b0 is the number of calories when slope is 0, b1 is the slope, x is either sandwhich or bistro box. We can use the mean to estimate the how many calories one would have depending on the type*

(c) Write down a hypothesis test for whether the mean calories for items in each group are the same or different.

> *Null Hypothesis: Calories are independent of whether it is a bistro box or sandwich Alt Hypothesis: Calories are dependent on the type of food Significance Level: 0.05*

(d) Fit your linear regression model for `calories` based on `type` to test whether there is a difference in mean calories between `"bistro box"` and `"sandwich"` items. Summarize your results using the `summary` function.

Hint: The syntax `lm(y ~ x)` will still work even if `x` is a binary explanatory variable.

```
# code you answer here
```

Based on the p-value results above and assuming an $\alpha = 0.05$ significance level, what would be the result of your previous hypothesis test?

> *REPLACE THIS TEXT WITH YOUR ANSWER*

(e) Instead of the linear regression approach above, now perform a **permutation test** to try and answer your 2-sample hypothesis test from earlier using $m = 1000$ repeats. Plot the resulting distribution of simulated test statistics using a histogram and then compute the corresponding 2-sided p-value.

Hint: Some of your code from HW4 might be helpful here.

```
set.seed(130)

# code you answer here
```

How does this p-value compare to the one computed using the linear regression-based test? Does your original conclusions (accept/reject) change as a result? Based on the number of observations in each group, in 1-2 sentences comment on which test (if any) you would consider more reliable and why.

> *REPLACE THIS TEXT WITH YOUR ANSWER*