STA130 Rstudio Homework

Problem Set 6

[Xuanqi Wei] ([1009353209]), 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)
```

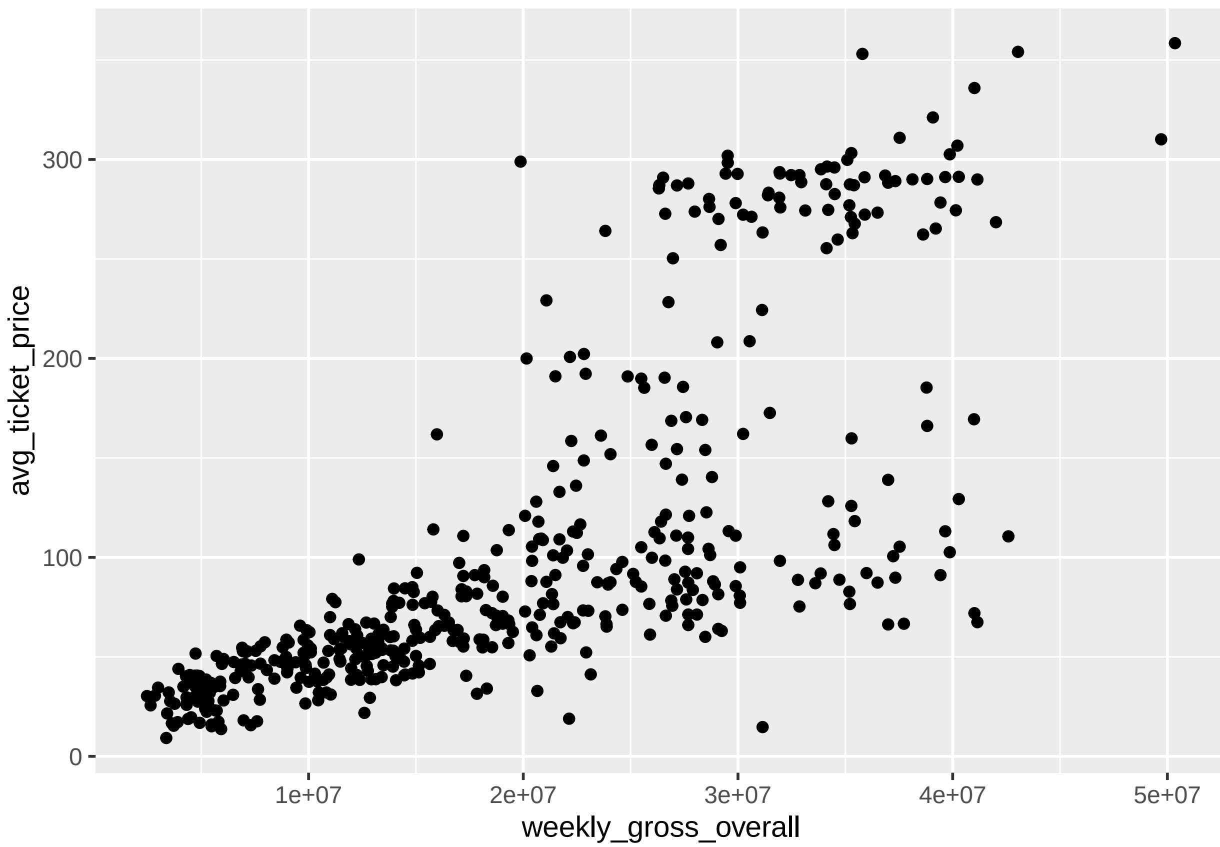
```
## Rows: 500
## Columns: 8
## $ show                   <chr> "La Cage aux Folles", "42nd Street", "42nd Street~
## $ Hamilton               <chr> "No", "No", "No", "No", "No", "No", "No", "No", "~
## $ week_ending            <date> 1985-07-28, 1985-09-08, 1985-09-15, 1985-12-15, ~
```

```
## $ 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).

```
broadway_data %>% ggplot(aes(weekly_gross_overall, 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.

> No, it's not appropriate to characterize and summarize the association inthe above plot with a striaight since the trend of this graph are two seperate parts.

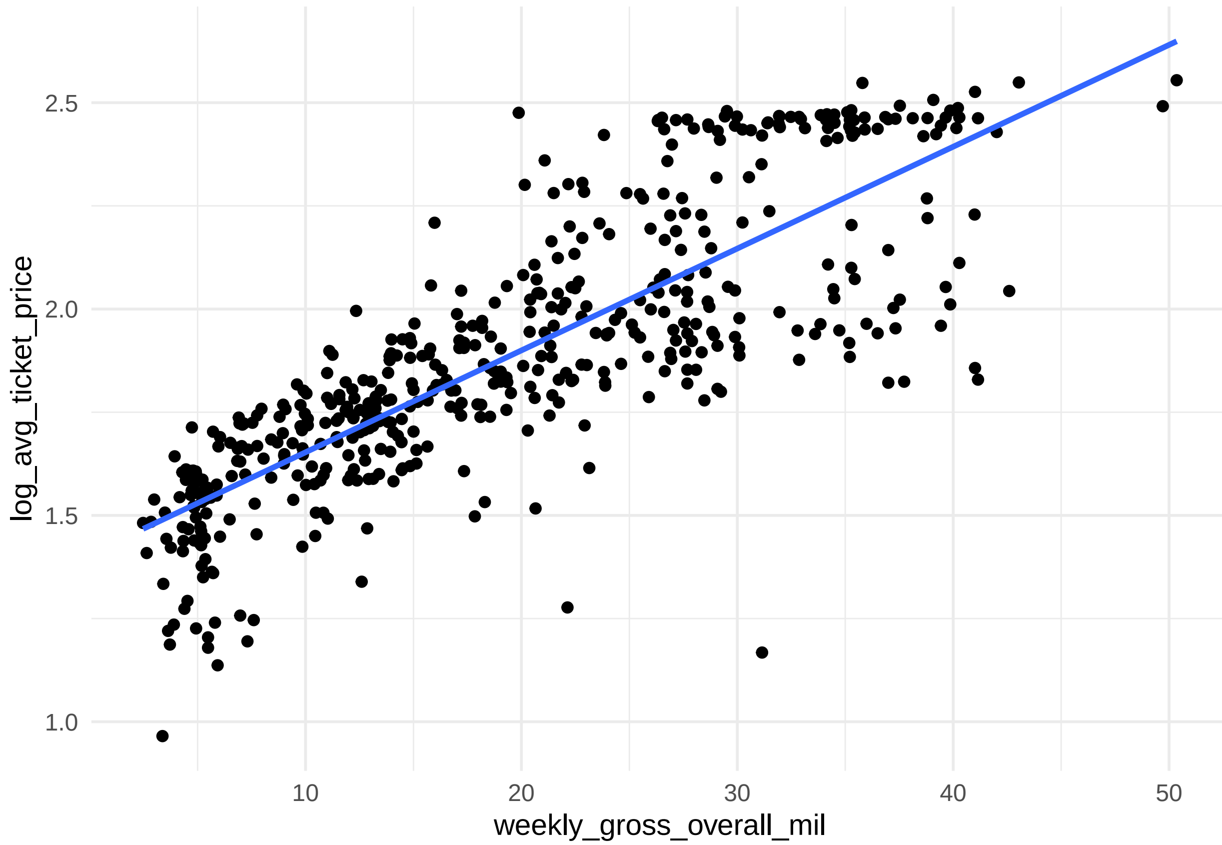
(b) Use the `mutate()` function to add the new variables `log_avg_ticket_price = log10(avg_ticket_price)` and `weekly_gross_overall_mil=weekly_gross_overall/1e6` to the data set.

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.*

```
broadway_data <- broadway_data %>% mutate(log_avg_ticket_price = log10(avg_ticket_price), weekly_gross_
```

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.

```
broadway_data %>%
ggplot(aes(weekly_gross_overall_mil, log_avg_ticket_price)) +
geom_point() +
geom_smooth(se=FALSE, method="lm") + theme_minimal()
```



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.

> *The transformation was helpful because the relationship between the weekly_gross_overall_mil and log_avg_ticket_price is a positive linear relationship.*

(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`.

```
correlation <- cor(broadway_data$log_avg_ticket_price,broadway_data$weekly_gross_overall_mil)
correlation
```

```
## [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.

> *These number implies* `log_avg_ticket_price` *and* `weekly_gross_overall_mil` *are strongly*

correlated.

(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}).

- $$y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

Now explain each component of the model above.

- y_i is the log10 of average ticket price.
- β_0 is the average of average ticket price while weekly gross overall mil equals to zero
- β_1 is the average change in average ticket price for 1 unit change in weekly gross overall mil.
- x_i is the weekly gross overall mil.
- $i = 1, \ldots, n$ where n is the number of show in the sample.
- ϵ_i is the random error term.

(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`.

- H_0: There is a linear relationship between the `weekly_gross_overall_100k` and `log_avg_ticket_price`.
- H_1: There is not a linear relationship between `weekly_gross_overall_100k` and `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.

```
b_f <- lm(log_avg_ticket_price ~ weekly_gross_overall_mil, data = broadway_data)
summary(b_f)$coefficients
```

```
##                           Estimate    Std. Error  t value      Pr(>|t|)
## (Intercept)              1.4061444 0.0175589041 80.08156 1.357759e-286
## weekly_gross_overall_mil 0.0246794 0.0007850831 31.43539 2.542862e-120
```

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.

> -To interpret the different rows/columns/entries from the `summary()` output in the context of the underlying data and model: - The average of average ticket pice when weekly gross overall mil equal to zero is:$\hat{\beta}_0 = 1.4061444$ - The average change in average ticket price for 1 unit change in weekly gross overall mil is: $\hat{\beta}_1 = 0.0246794$ - To reject H_0: the p-value < 2e-16

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

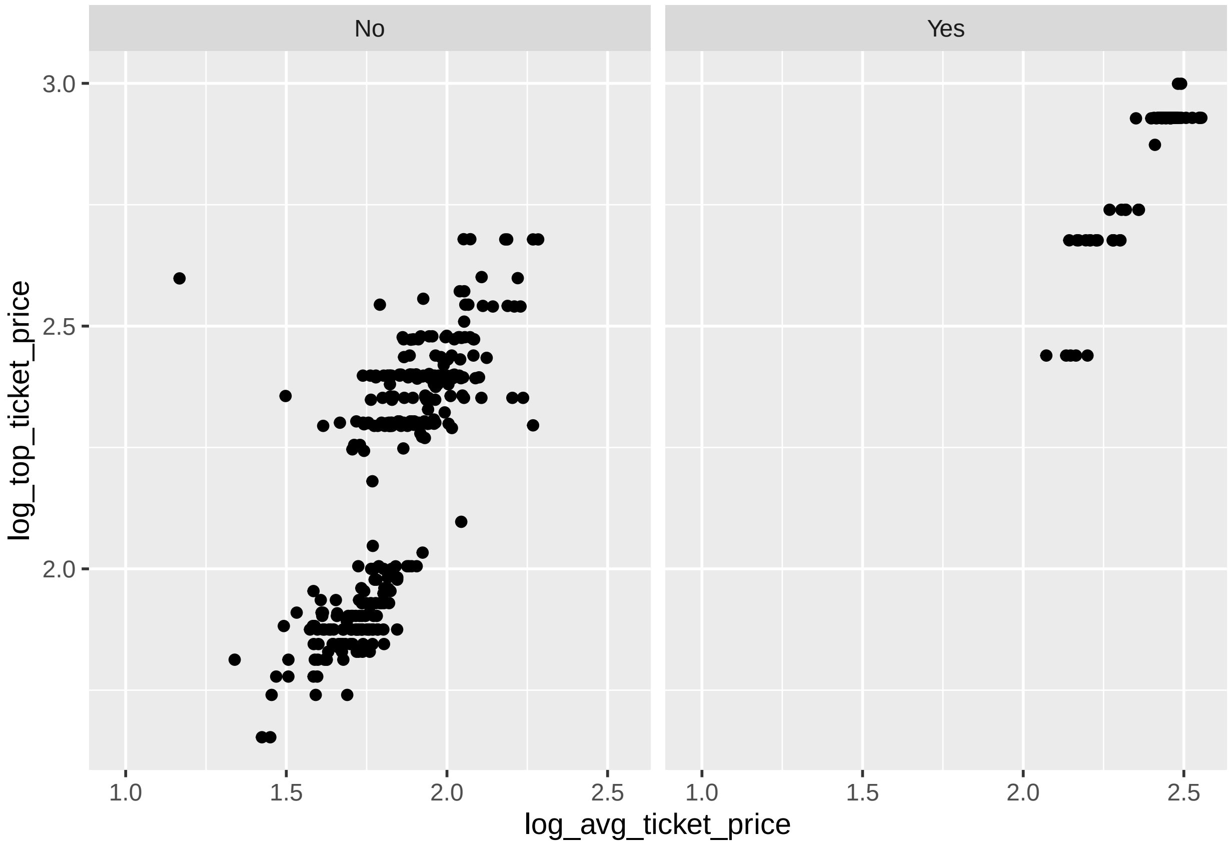
- To draw a conclusion regarding the hypothesis test, $\alpha = 10^{-3} >$ p-value which is less than 2e-16.

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.

```
broadway_data <- broadway_data %>% mutate(log_top_ticket_price = log10(top_ticket_price))
broadway_data %>% ggplot(aes(log_avg_ticket_price, log_top_ticket_price)) + geom_point() + facet_wrap(~H
```



(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.

```
broadway_data %>% filter(!is.na(log_top_ticket_price)) %>% group_by(Hamilton) %>%
summarise(correlation = cor(log_avg_ticket_price, log_top_ticket_price))
```

```
## # A tibble: 2 x 2
##   Hamilton correlation
```

```
##   <chr>          <dbl>
## 1 No             0.757
## 2 Yes            0.929
```

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.

> *To discuss the correlations I computed above, there are not any big differences between whether the musical was Hamilton or not since the number of Yes is slightly biggere tha. the number of No. Thus,`log_top_ticket_price` and `log_avg_ticket_price` weakly relates to each other.*

(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.

```
b_f_d_1 <- broadway_data %>% filter(Hamilton == "Yes")
b_f_d_2 <- broadway_data %>% filter(Hamilton == "No")
b_f_1 <- lm(log_top_ticket_price ~ log_avg_ticket_price, data = b_f_d_1)
b_f_2 <- lm(log_top_ticket_price ~ log_avg_ticket_price, data = b_f_d_2)
summary(b_f_1)
```

```
##
## Call:
## lm(formula = log_top_ticket_price ~ log_avg_ticket_price, data = b_f_d_1)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.181450 -0.018282  0.001623  0.029456  0.128579
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)           0.01697    0.11413   0.149    0.882
## log_avg_ticket_price  1.18350    0.04753  24.899   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.05075 on 98 degrees of freedom
## Multiple R-squared:  0.8635, Adjusted R-squared:  0.8621
## F-statistic:   620 on 1 and 98 DF,  p-value: < 2.2e-16
```

```
summary(b_f_2)
```

```
##
## Call:
## lm(formula = log_top_ticket_price ~ log_avg_ticket_price, data = b_f_d_2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.3876 -0.1293 -0.0010  0.1051  1.1890
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)           0.05752    0.10577   0.544    0.587
## log_avg_ticket_price  1.15768    0.05723  20.229   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

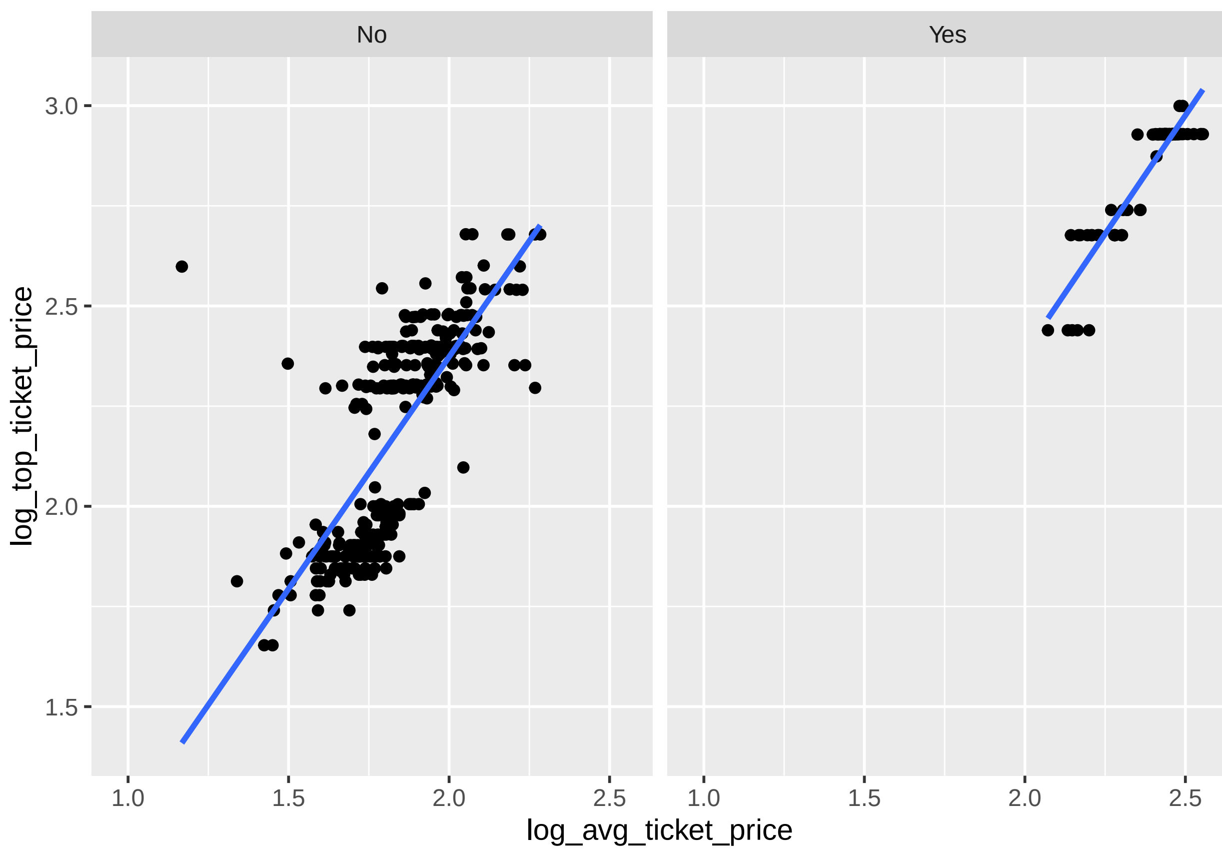
```
##
## Residual standard error: 0.1708 on 304 degrees of freedom
##   (94 observations deleted due to missingness)
## Multiple R-squared:  0.5738, Adjusted R-squared:  0.5724
## F-statistic: 409.2 on 1 and 304 DF,  p-value: < 2.2e-16
```

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?

> -To comment on what the fitted coefficients (slope and intercept) of the model implies for the relationship between `log_top_ticket_price` and `log_avg_ticket_price`, firstly, for Hamilton Musical: The log_top_ticket_price when log_ave_ticket_price equal to zero, the $\hat{\beta}_0 = 0.01697$; The log_top_ticket_price for 1 unit change in log_ave_ticket_price, the $\hat{\beta}_1 = 1.18350$. To reject H_0, p-value $<$ 2e-16. - For non-Hailton Musical: The log_top_ticket_price when log_ave_ticket_price equal to zero, the $\hat{\beta}_0 = 0.05752$; The log_top_ticket_price for 1 unit change in log_ave_ticket_price, the $\hat{\beta}_1 = 1.15768$. To reject H_0, p-value $<$ 2e-16.

(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)`.

```
broadway_data <- broadway_data %>% mutate(log_top_ticket_price =log10(top_ticket_price))
broadway_data %>% ggplot(aes(log_avg_ticket_price, log_top_ticket_price)) +
geom_point() + geom_smooth(method=lm, se=FALSE) + facet_wrap(~Hamilton)
```



Question 3: Starbucks

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

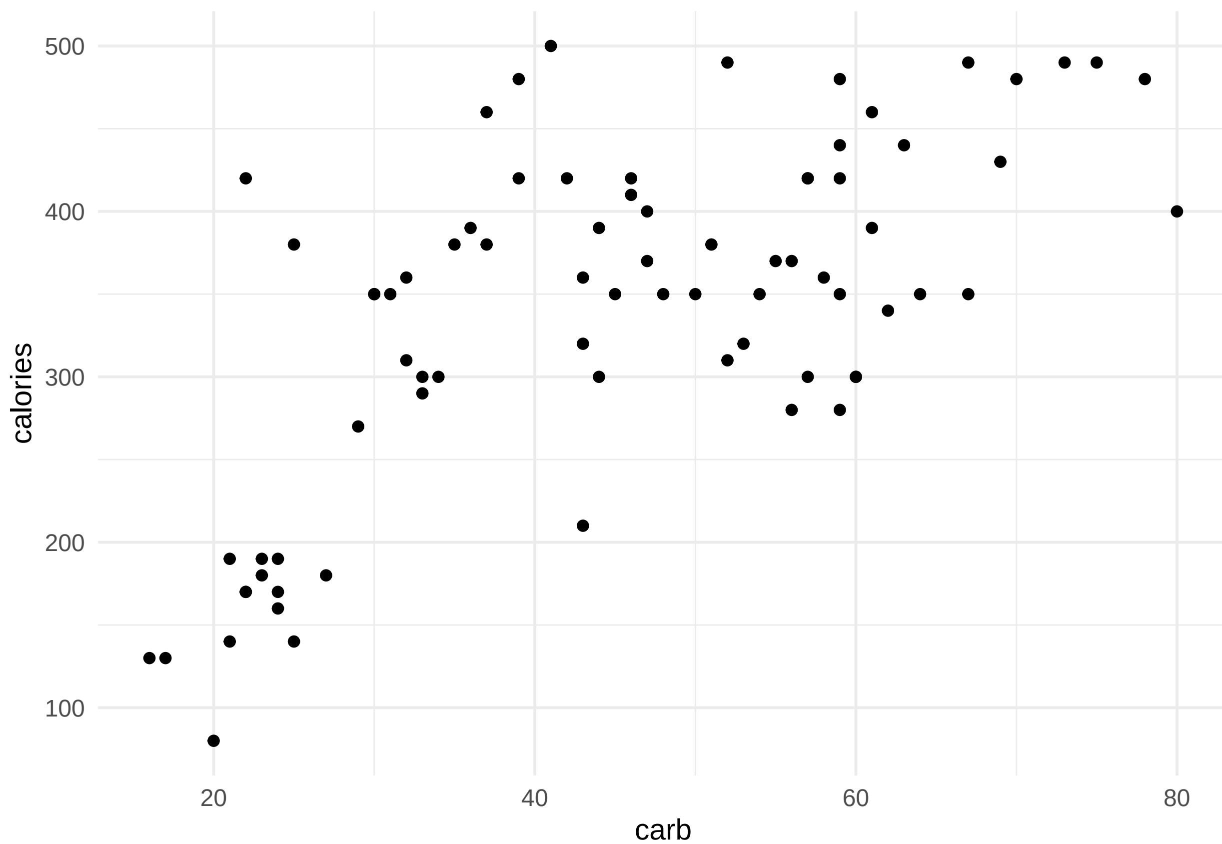
```r
# load in data
starbucks_data <- read_csv("starbucks.csv")

# 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.

```r
starbucks_data %>% ggplot(aes(carb, calories)) + geom_point() + theme_minimal()
```



Write 1-2 sentences describing any association you observe.

The `carb` and `calories` have a positie linear association.

(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.

Since the association is positive and is relatively moderate, the correlation coefficient between carbohydrates and calorie content in Starbucks menu might be 0.7.

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

```
sb_correlation <- cor(starbucks_data$carb,starbucks_data$calories)
sb_correlation
```

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

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

Similar.

(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.

```
l_r_m <- lm(calories ~ carb, data = starbucks_data)
l_r_m_sum <- summary(l_r_m)
l_r_m_sum
```

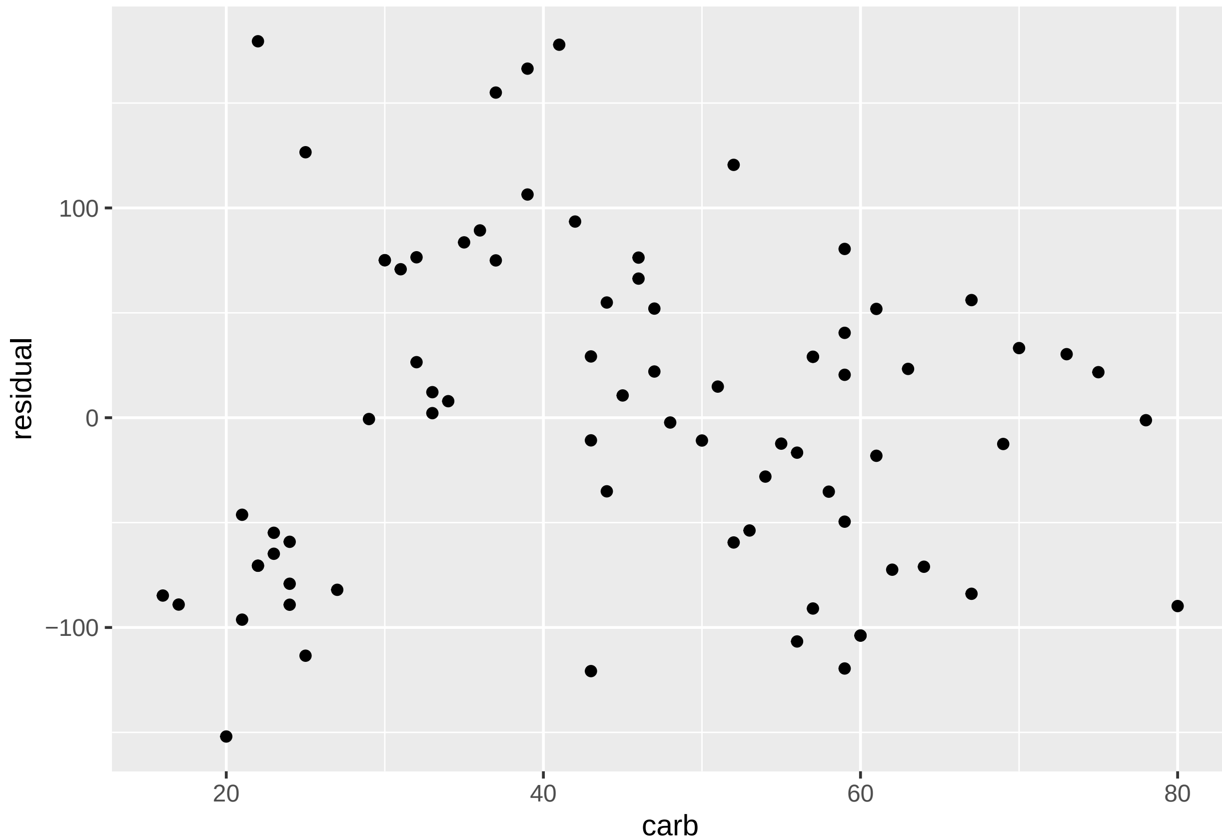
```
##
## Call:
## lm(formula = calories ~ carb, data = starbucks_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -151.962  -70.556   -0.636   54.908  179.444
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 146.0204    25.9186   5.634 2.93e-07 ***
## 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
```

the main results highlighted in the `summary()` output are the calories when carb equal to zero which $\hat{\beta}_0 = 146.0204$; the calories for 1 unit change in carb which $\hat{\beta}_1 = 4.2971$. To reject H_0, the p-value is 1.67e-11.

(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.

```
s_residual <- l_r_m$residuals
tibble(residual = s_residual, carb = starbucks_data$carb) %>% ggplot(aes(carb, residual)) + geom_point(
```



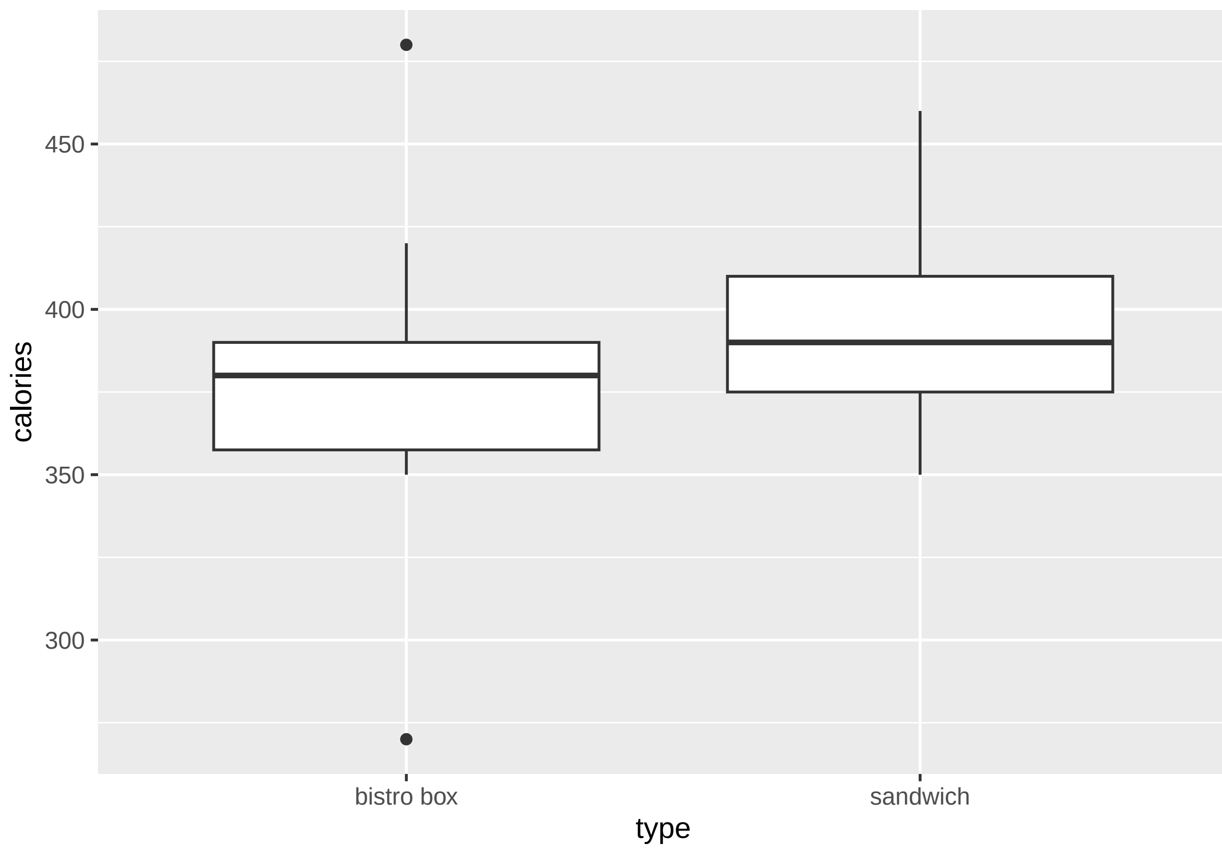
In 1-2 sentences, comment on any trends (or lack of trends) that you may observe and what this implies about the overall fitted relationship.

No, it doesn't show a linear relationship.

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



(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.

-

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

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

- The average calories for bistro box is $\hat{\beta}_0$
- The different in average calories between bistro box and sandwich is $\hat{\beta}_1$

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

- Null Hypothesis H_0: $\beta_0 = 0$
- H_1: $\beta_0 \neq 0$

(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.

```
b_f <- lm(calories ~ type, data = starbucks_lunch)
summary(b_f)
```

```
##
## Call:
## lm(formula = calories ~ type, data = starbucks_lunch)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -107.50  -22.50    2.50   14.29  102.50
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)    377.50      17.85  21.153 1.87e-11 ***
## typesandwich    18.21      26.12   0.697    0.498
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 50.48 on 13 degrees of freedom
## Multiple R-squared:  0.03605,    Adjusted R-squared:  -0.0381
## F-statistic: 0.4861 on 1 and 13 DF,  p-value: 0.4979
```

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?

- Based on the p-value results above and assuming an $\alpha = 0.05$ significance level, we have no evidence to reject H_0 since $p - value = 0.498$.

(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.

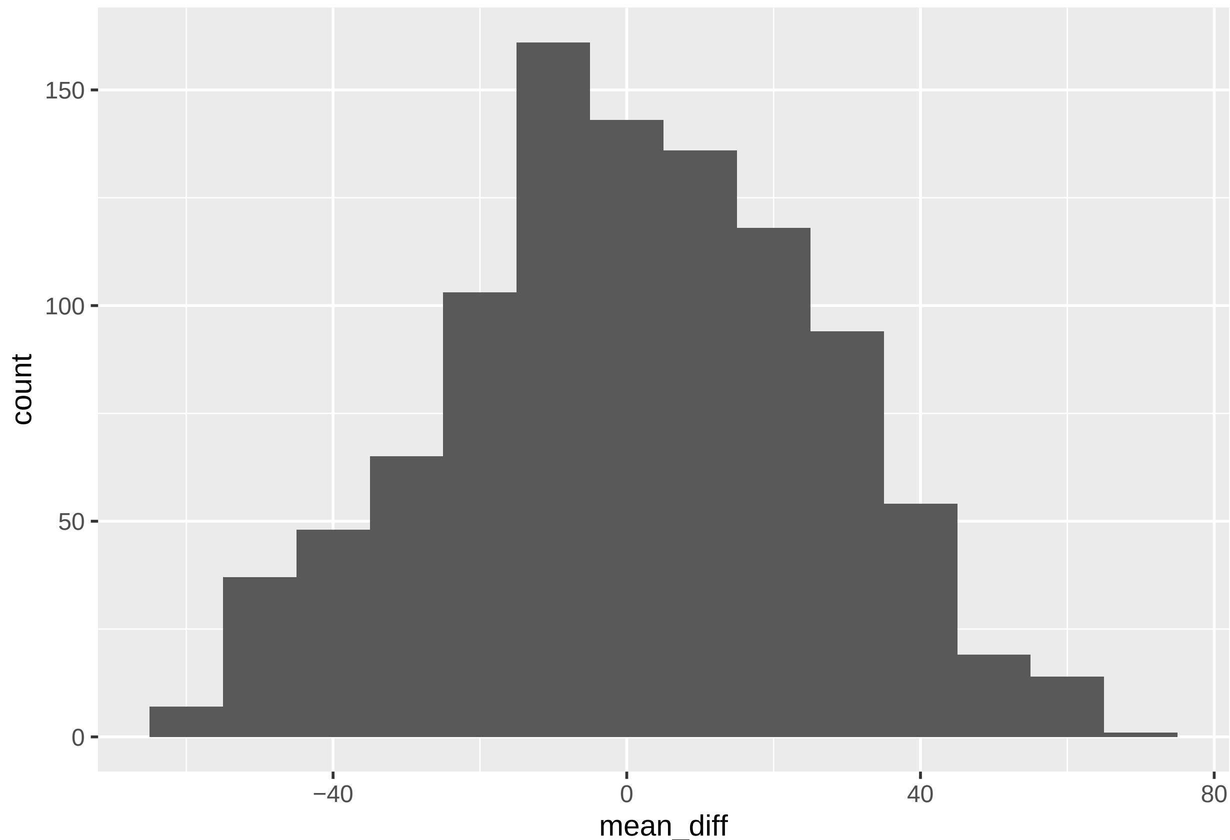
```
set.seed(130)

t_stat <- starbucks_lunch %>% group_by(type) %>%
summarise(means = mean(calories), .groups="drop") %>%
summarise(value = diff(means))
t_stat <- as.numeric(t_stat)
t_stat
```

```
## [1] 18.21429
```

```
repetitions <- 1000;
simulated_values <- rep(NA, repetitions)
for(i in 1:repetitions){
simdata <- starbucks_lunch %>% mutate(type = sample(type))
sim_value <- simdata %>% group_by(type) %>%
summarise(means = mean(calories), .groups="drop") %>%
summarise(value = diff(means))
```

```
simulated_values[i] <- as.numeric(sim_value)
}
sim <- tibble(mean_diff = simulated_values)
sim %>% ggplot(aes(x=mean_diff)) + geom_histogram(binwidth=10)
```




```
num_more_extreme <- sim %>% filter(abs(mean_diff) >= abs(t_stat)) %>% summarise(n())
p_v <- as.numeric(num_more_extreme / repetitions)
p_v
```

```
## [1] 0.492
```

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.

- This p-value compare to the one computed using the linear regression-based test is smaller. The original conclusions doesn't change as we still support H_0. Based on the number of observations in each group, the linear regression-based test is more reliable, since our result is based on sample.