More about Your Starbucks Order

Liuyixin Shao, Sophia Lan, Cole Smidt

2022-03-13

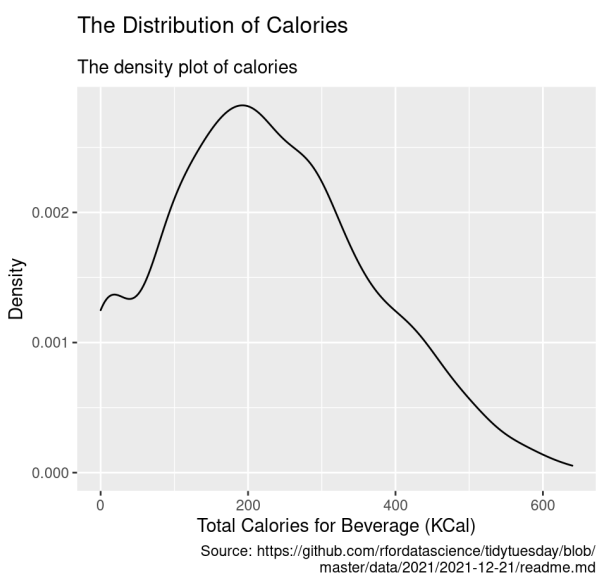
Introduction

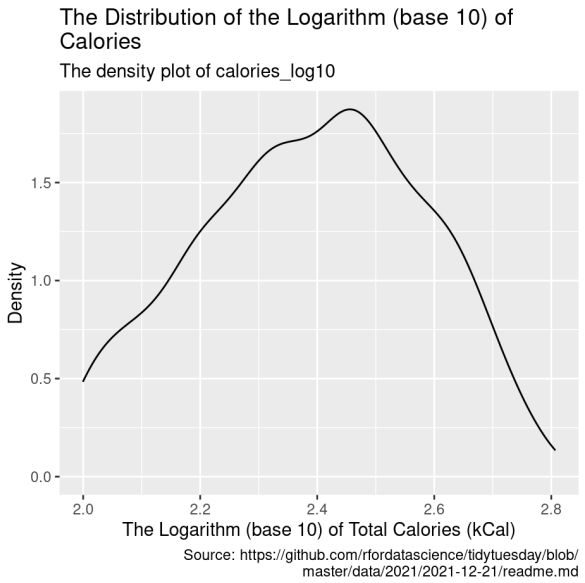
How many calories are in your go-to Starbucks order? How does this calorie count change from drink to drink, and what happens if I get a venti iced latte instead of a grande, do the extra calories really add up? What variables influence the amount of calories in a given drink the most?

Starbucks is the largest coffee shop in the world, selling upwards of 4 billion cups of coffee every year according to their website. Millions of people across the world, probably including you, go to Starbucks every morning for their daily coffee. The Official Starbucks Nutritional dataset (see link 1), released by Starbucks, provides customers with information on all beverages offered on their menu, but is difficult to interpret with just a glance.

The interpretation of this information can help customers understand the nutritional content and calorie profile of their favorite beverages. First of all, calorie count in individual servings can help us better assess the total calories we consume in a day. Secondly, some nutrients may help some special groups make choices that can be vital to their health. For example, people with diabetes will choose beverages that are sugar-free, lower in fat, and lower in calories, and people who are lactose intolerant can choose milk-free drinks. Customers can find this relevant information from this data sheet and the interpretations made from it. Third, if a very detailed nutritional data table is provided, customers will trust Starbucks more, which can improve the reputation of the brand and attract more customers. Collecting this data is also very useful for Starbucks in making its menu. The menu contains guides on how customers can make better choices based on calories and other factors.

As seen in the graph below, the distribution of calories has a fairly significant right-skew. It is also somewhat bimodal, with a larger distribution near zero. In order to combat this and make the data more normal, we have converted our response variable `calories` to a logarithmic scale (base 10), named `calories_log10` , thus making it a more normal and less skewed distribution, as seen below. Along with this, we also cut the dataset to no longer include any drinks with less than 100 calories to eliminate outliers in our dataset. Our new response variable, also below, has a much more normal distribution, making it much easier to model the data.





Then, we converted milk from an integer variable to a categorical variable, making it easier to use with visuals and model building. We also separate “venti” in `size` into two different categories - “venti(cold)” and “venti(hot)”, since “venti” size cups for cold drinks and hot drinks in Starbucks contain different serving sizes.

Below is a codebook of the dataset that we used, along with a glimpse of our data as well.

Data Name	Description	Class	Values
`product_name`	Name of the product	categorical	Coffee, Latte, Espresso, etc
`size`	Size of drink	categorical	short, tall, grande, venti(hot), venti(cold), trenta
`milk`	Type of milk used in drink	categorical	none, nonfat, 2%, soy, coconut, whole
`whip`	Whether or not whipped cream was added	categorical	yes, no
`calories`	Kilocalorie in product	numeric	$[100, 640]$
`calories_log10`	The logarithm of total calories in product (kCal)	numeric	$[2, 2.80618]$
`total_fat_g`	Total fat in grams	numeric	$[0, 28]$
`sodium_mg`	Sodium in milligrams	numeric	$[0, 370]$

Data Name	Description	Class	Values
total_carbs_g	Total Carbohydrates in grams	numeric	$[8, 96]$
sugar_g	Sugar in grams	numeric	$[5, 89]$

```
## Rows: 947
## Columns: 10
## $ product_name   <chr> "Caffè Misto", "Caffè Misto", "Caffè Misto", "Caffè Mis…
## $ size           <chr> "tall", "grande", "grande", "venti(hot)", "venti(hot)",…
## $ milk           <chr> "whole", "2%", "whole", "2%", "whole", "soy", "none", "…
## $ whip           <chr> "no", "no", "no", "no", "no", "no", "no", "no", "no", "…
## $ calories       <dbl> 100, 110, 130, 130, 160, 100, 120, 160, 100, 110, 120, …
## $ calories_log10 <dbl> 2.000000, 2.041393, 2.113943, 2.113943, 2.204120, 2.000…
## $ total_fat_g    <dbl> 5.0, 4.0, 7.0, 5.0, 9.0, 4.0, 0.1, 0.1, 0.1, 1.5, 2.0, …
## $ sodium_mg      <dbl> 70, 100, 95, 130, 120, 100, 15, 15, 35, 40, 35, 35, 35,…
## $ total_carbs_g  <dbl> 8, 10, 10, 13, 13, 8, 30, 40, 23, 23, 23, 22, 22, 35, 3…
## $ sugar_g        <dbl> 8, 10, 10, 13, 13, 6, 30, 39, 23, 23, 23, 21, 22, 35, 3…
```

Model Building

We have built three separate models to try to estimate the number of calories in a given drink. In order to do this, we split the dataset in half into a training group and a testing group. We have built the models on the training data and will test their validity of the testing set to see which of the three is able to model the best.

Model proposed by Liuyixin Shao

I chose to use `total_carbs_g` , `total_fat_g` , and `size` as the explanatory variables in my model. The fitted model equation I came up with for the response variable is:

$$\hat{calories\_log10} = 1.99 - 0.0277 * sizeshort - 0.0106 * sizetall - 0.0295 * sizetrenta - 0.0289 * sizeventi(cold)$$
$$+0.0153 * sizeventi(hot) + 0.0067 * total\_carbs\_g + 0.0148 * total\_fat\_g$$

This means, holding the other variables in the model the same, when the total carbohydrate in the beverage increases by 1 gram, the calories it contains is expected to increase by a factor of 1.0155, on average. When the total fat increases by 1 gram, the calories the beverage contains is expected to increase by a factor of 1.0347, on average. Holding the other variables in the model the same, if a customer orders a drink with “short” size, the calories inside will be expected to decrease by a factor of 0.9382, on average, while “tall” size to decrease by a factor of 0.9759, “trenta” size to decrease by a factor of 0.9343, “venti(cold)” size to decrease by a factor of 0.9356, and “venti(hot)” size to increase by a factor of 1.0359, on average.

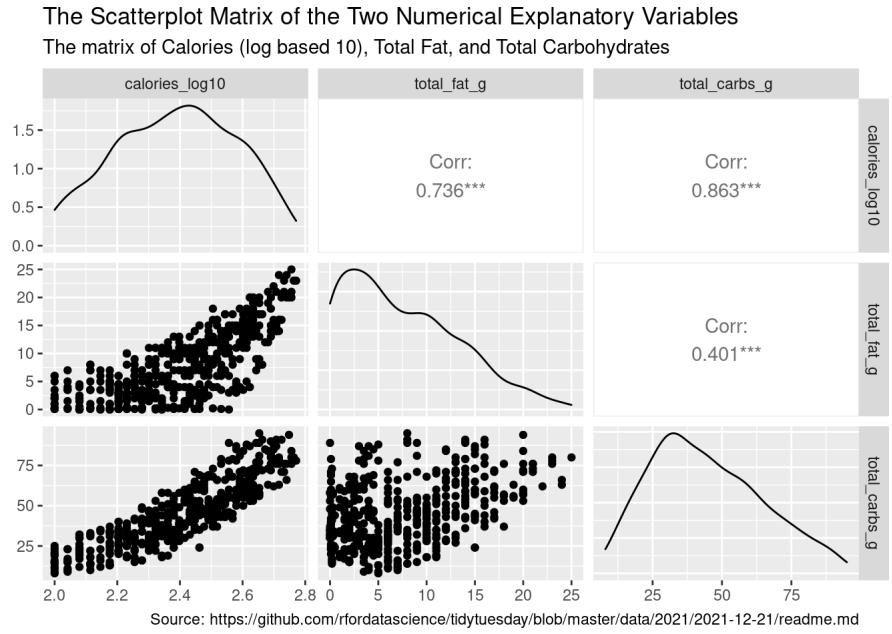
I put all the estimates from the “tidy” function as well as the r-square and RMSE I calculated below.

```
## # A tibble: 8 × 3
##   term            estimate undo_log
##   <chr>              <dbl>    <dbl>
## 1 (Intercept)      1.99      97.5
## 2 sizeshort       -0.0277     0.938
## 3 sizetall        -0.0106     0.976
## 4 sizetrenta      -0.0295     0.934
## 5 sizeventi(cold) -0.0289     0.936
## 6 sizeventi(hot)   0.0153     1.04
## 7 total_carbs_g    0.00670    1.02
## 8 total_fat_g      0.0148     1.03
```

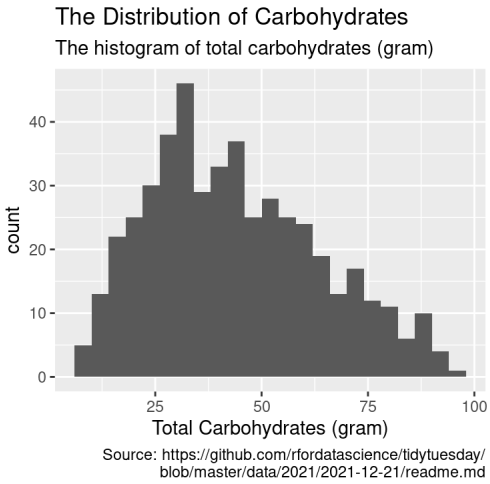
```
##      r.squared  adj.r.squared rmse(standard)
##     0.93164620     0.93061720     0.05284181
```

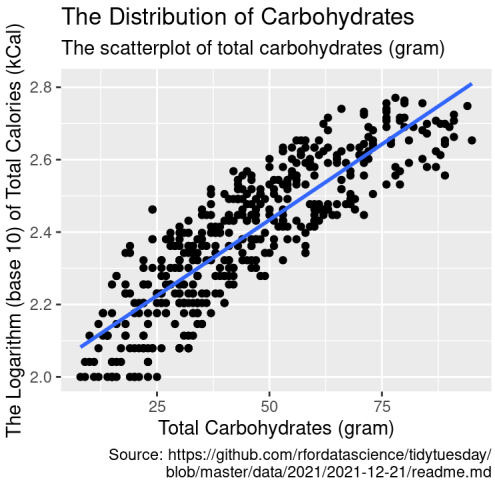
As we know, the three main categories of factors that affect calories in food are carbohydrates, fat, and protein. Therefore, I’d like to pick my explanatory variables that are related to these categories.

Since `total_fat_g` and `total_carbs_g` are two existing variables in our Starbucks dataset, I started my analysis by pre-selecting them as two of my explanatory variables. Then, I made a scatterplot matrix to roughly see the distributions of my two pre-selecting numerical variables. According to the matrix below, both `total_fat_g` and `total_carbs_g` show strong linear positive relationships with `calories_log10` , which is our response variable. The correlation is about 0.736 and 0.863 separately, showing that they are highly correlated with the response variable. What’s more, the correlation between these two numerical variables is 0.401, which is weak and proves that I didn’t encounter collinearity for choosing these two as my explanatory variables.

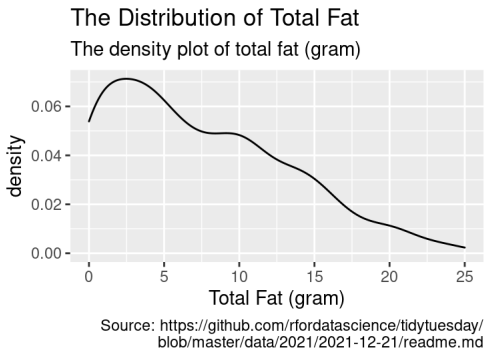


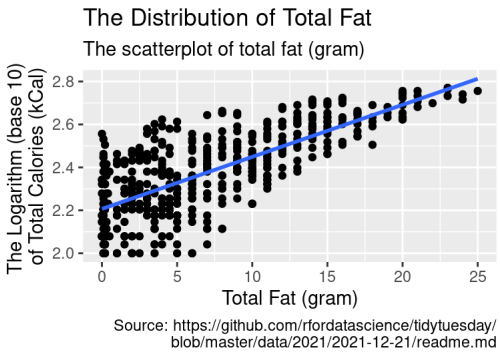
Let's take a closer look at the distribution of `total_carbs_g`. Even though the distribution in the histogram is slightly right-skewed, I decided not to take logarithm to this variable because 1) the distribution will become left-skewed when I apply log in `total_carbs_g`, 2) there exists a strong linear relationship in the scatterplot of `total_carbs_g` and `calories_log10`. With further analysis, I knew the median of `total_carbs_g` in this training dataset is 42 grams with a range from 8 to 95 grams, verifying that the distribution of `total_carbs_g` is roughly symmetric. The correlation between `total_carbs_g` and `calories` is above 87%, proving that the relationship between these two variables is very strong. Thus, I picked `total_carbs_g` as one of my numerical explanatory variables.

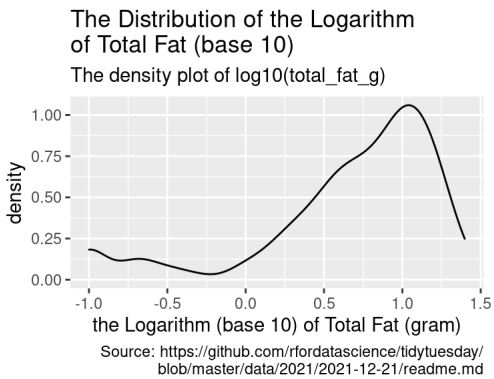


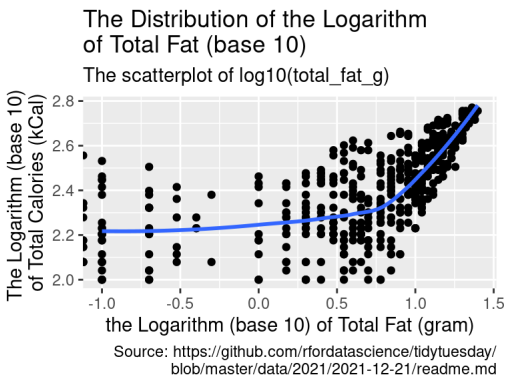


Below are the distributions of `total_fat_g`. Although the distribution in the histogram is also right-skewed and taking logarithm to `total_fat_g` will make the distribution of `total_fat_g` slightly more symmetric, I decided not to take logarithm to this variable because 1) the distribution in scatterplot will become nonlinear when I apply log in `total_fat_g`, 2) there exists a strong linear relationship in the scatterplot of `total_fat_g` and `calories_log10`. I put the difference between the distribution of `total_fat_g` versus the `calories_log10` and the logarithm version of `total_fat_g` versus the `calories_log10` below.







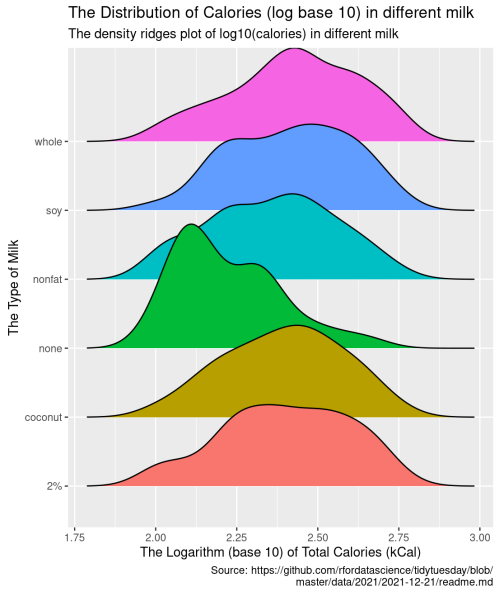


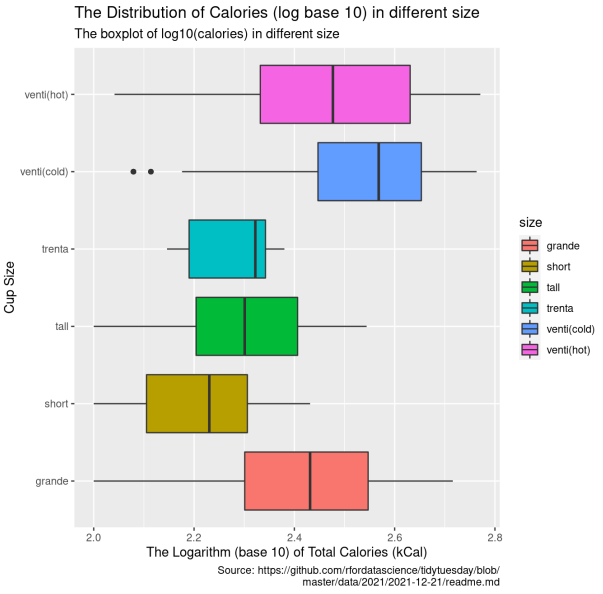
Then I used “bootstrap” to further prove the strong positive relationship between total fat and calories in Starbucks beverages. Because taking the logarithm of calories made my response variable small in scale, resulting in the confidence interval of the slope I calculated between `calories_log10` and `total_fat_g` very small (but still meaningful because the range doesn’t contain zero), I chose to calculate my confidence interval with the actual calories in the dataset. In this way, even though I’m using the training dataset and don’t know what the full population looks like, I’m 95% confident that the slope of `total_fat_g` in the population is between 14.389 and 16.392. Since 0 is not in this range, there must be a positive relationship between total fat and calories. Therefore, I chose `total_fat_g` as one of my explanatory variables.

```
## lower ci for log10_calories upper ci for log10_calories
##                       0.023                       0.026
##       lower ci for calories       upper ci for calories
##                      14.389                      16.392
```

(Notice: “ci” represents the word “confidence interval”)

Since protein rarely exists in beverage drinks, there isn’t a column that specifically analyzes protein in our Starbucks dataset. I thought `milk` would be the most appropriate variable that is relevant to protein. However, according to the density plot I made below, there isn’t a huge difference in the distributions between different kinds of milk and calories (except for “non-milk”, which is undoubtedly reasonable to see this difference because beverages without milk also contain less fat, and we already knew fat has a strong correlation with calories). As a result, I gave up picking `milk` as one of my explanatory variables because the correlation between `milk` and `calories` is weak.





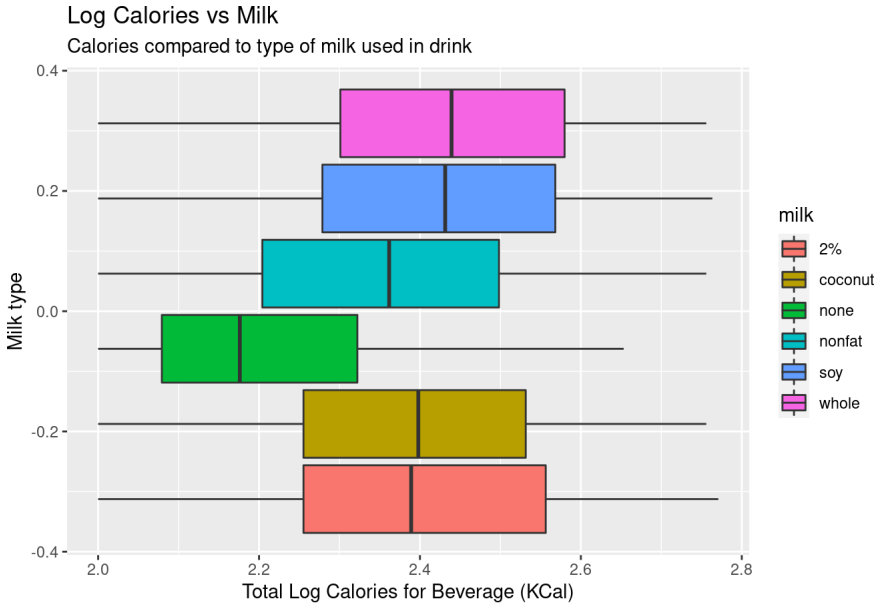
The size of beverage could also be a factor that affects its calories inside, as we know taking other relevant factors the same, a larger cup of Starbucks drink always contains more calories than a smaller one. According to the boxplot above and the table below, the median amount of calories in the “short” size cups (which are the smallest serving cups in Starbucks) is 170 kCal. “Tall” and “grande” are the second and the third smallest serving cups in Starbucks. Their median amounts of calories are 200 kCal and 270 kCal separately. The serving size of “venti” cups is 709 ml for cold drinks and 591 ml for hot drinks. Therefore, as we expected, the median amount of calories in “venti” size cups for cold drinks is 370 kCal, which is larger than those for hot drinks (which is around 300 kCal). Surprisingly, “trenta” is the largest cup size served in Starbucks, but the median amount of calories it contains is only 210 Kcal. This is because “trenta” size cups are only available in specific iced beverages like iced coffee, cold brew, and tea drinks with more ice but not flavored drinks inside (see link 2). The seemingly anomalous but reasonably low calories content in “trenta” size Starbucks beverages also becomes the reason why I chose to use the categorical variable `size` instead of the numerical variable `serv_size_m_l` in the dataset for my model. Otherwise, my model won’t be very accurate. In summary, there is a distinct relationship between the categories of serving cup size and the logarithm of calories in Starbucks drinks, resulting in `size` is a suitable categorical explanatory variable for my model.

```
## # A tibble: 6 × 4
##   size        median_size IQR_size undo_log_median
##   <chr>             <dbl>    <dbl>           <dbl>
## 1 grande             2.43    0.246            270
## 2 short              2.23    0.201            170
## 3 tall               2.30    0.202            200
## 4 trenta             2.32    0.152            210
## 5 venti(cold)        2.57    0.206            370
## 6 venti(hot)         2.48    0.299            300.
```

Model Proposed by Cole Smidt

In the model that I have created, I will be comparing the number of calories to the milk type used, the total amount of fat a drink has, the amount of sodium a drink has, and the amount of sugar in a given drink.

The first variable used in my model, `milk`, is fairly self-explanatory. There are different amounts of fat in each type of milk, and because fat directly plays into the total calorie count, with each gram of fat adding 9 calories to the total calorie count. As seen with the graph below, there is certainly a correlation between milk and calories.

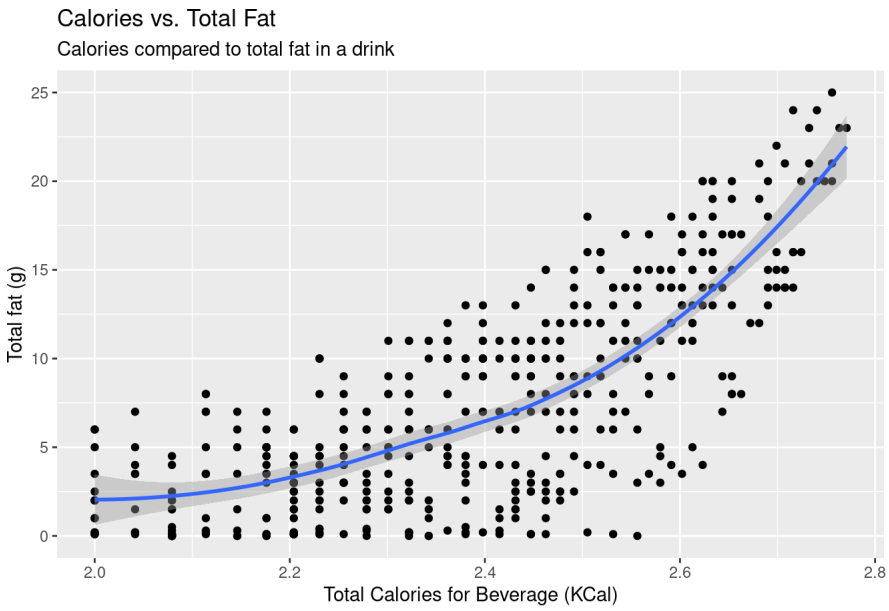


Source: https://github.com/rfordatascience/tidytuesday/blob/master/data/2021/2021-12-21/readme.md

```
## # A tibble: 6 × 4
##   milk    median_milk IQR_milk undo_log_median
##   <chr>         <dbl>    <dbl>           <dbl>
## 1 2%             2.39    0.301            245.
## 2 coconut        2.40    0.276            250
## 3 none           2.18    0.243            150
## 4 nonfat         2.36    0.294            230
## 5 soy            2.43    0.289            270
## 6 whole          2.44    0.279            275.
```

The results of the graph above were somewhat surprising to me until I conducted a little bit more research. I was intrigued by the high average of coconut milk, until I learned that Starbucks adds sugar into its coconut milk, which is the reason that the average calorie count for drinks with coconut milk is statistically higher than that of nonfat and soy milk.

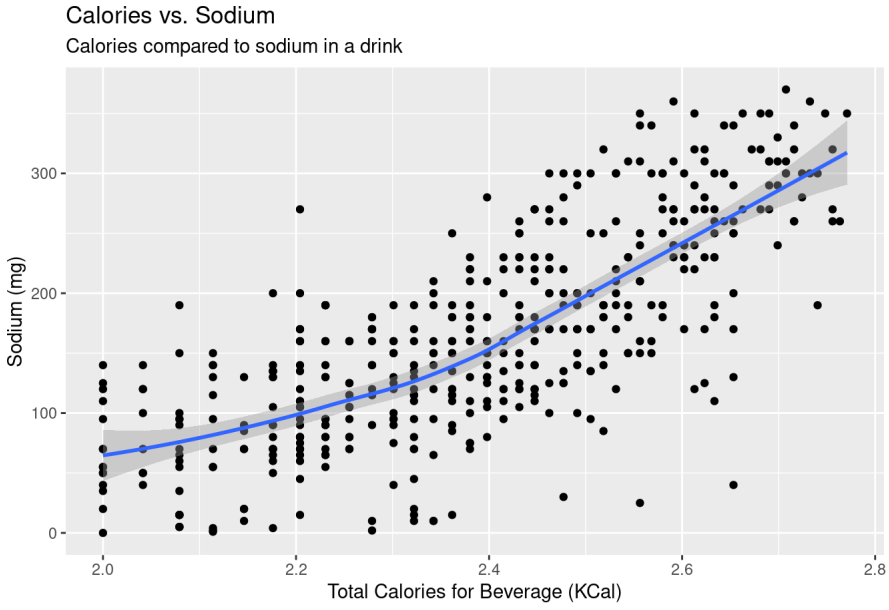
The next variable, `total_fat_g`, is also very intuitive when it comes to the comparison between it and calories. The amount of fat in a drink directly effects the total number of calories in a drink, which means that calculating the number of calories by fat will make a solid model. As seen by the graph below, there is a strong correlation between the two.




```
## # A tibble: 1 × 2
##    lower  upper
##    <dbl>  <dbl>
## 1 0.0227 0.0260
```

As seen by the graph, there is a strong correlation between the two variables. Along with this, I computed the 95 percentile range for the bootstrap of this variable, and found that it did not include zero, meaning that there is a strong positive correlation.

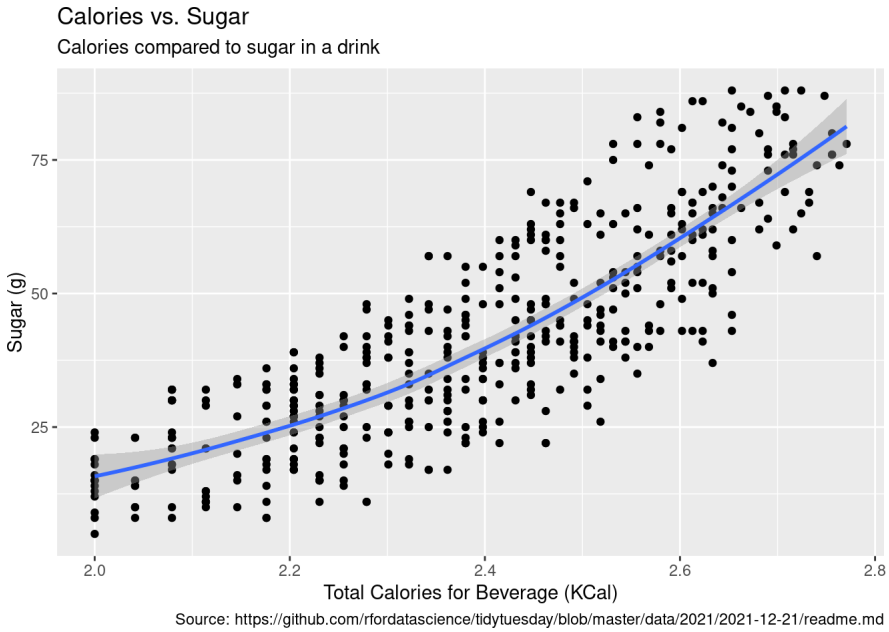
The third variable that will be a part of my model is `sodium_mg`. Sodium is not something inherently found in coffee, but is often found in other ingredients commonly added to specialty coffee drinks. It is also somewhat less intuitive to use this variable, as sodium does not directly effect the number of calories in a drink. It is useful, though, as it shows when other ingredients are added to the coffee, which often include things that are high in sugar and calories.




```
## # A tibble: 1 × 2
##     lower   upper
##     <dbl>   <dbl>
## 1 0.00157 0.00182
```

As seen above in the graph, there is a positive correlation between these two variables, and the 95% bootstrap distribution confirms this, with a lower bound of 0.0227 and an upper bound of 0.0260.

My fourth and final variable, which I found to be the best at modeling for calories, is `sugar_g`. I found this variable fairly intuitive to use here, as every gram of sugar in a drink adds 4 calories. Along with this, most calories from sweet coffee drinks comes from added sugar. Below you can see how well sugar models the number of calories just by itself.




```
## # A tibble: 1 × 2
##     lower   upper
##     <dbl>   <dbl>
## 1 0.00771 0.00862
```

The bootstraps for sugar has a confidence interval that are all positive, meaning there is a positive linear correlation between sugar and the amount of calories in a given drink. The linear correlation can also be seen by the graph.

With all this said, the model that I created has the following equation:

$$\hat{calories\_log10} = 2.01 - 0.0312 * milkcoconut - 0.0861 * milknone + 0.00757 * milknonfat + 0.0193 * milksoy * 0.00328 * milkwhole + 0.($$

When fitted to the test data, the RMSE I have calculated is below, along with the R squared and adjusted R squared values:

```
## # A tibble: 1 × 3
##   .metric .estimator .estimate
##   <chr>   <chr>          <dbl>
## 1 rmse    standard      0.0628
```

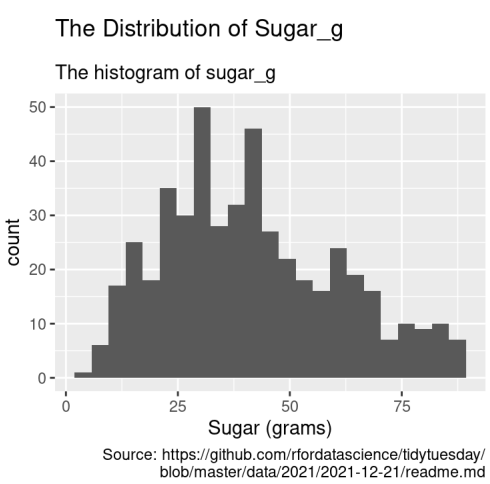
```
## # A tibble: 1 × 2
##   r.squared adj.r.squared
##       <dbl>         <dbl>
## 1     0.917         0.915
```

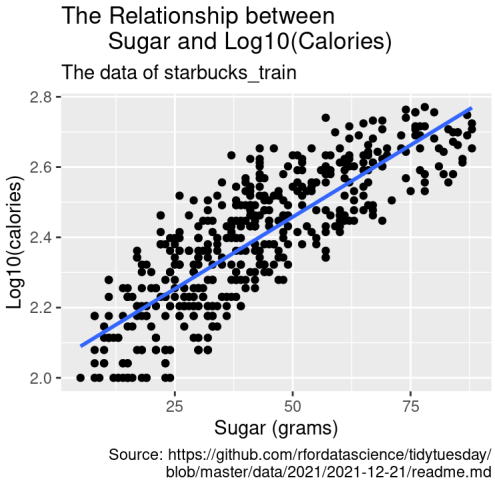
Model proposed by Sophia Lan

In my individual project, I am going to talk about the connection between 3 explanatory variables— `whip` , `sodium_mg` , and `sugar_g` —and our group's chosen response variable—Calories. Sodium and Sugar are the nutrients we see most often in our lives. Excessive salt can easily lead to increased blood pressure. Excessive sugar is more likely to cause obesity and diabetes. Whip is something that young people like to add to coffee very much. I'm predicting that these three explanatory variables may have a strong connection to calories. Below, I will use some statistical methods to prove my guess and build a Model.

For sugar, as the website Jessicablack says, I quote a paragraph - "Hand-in-hand with limiting sugar intake is avoiding empty calories. Empty calories are defined as calories that provide a quick burst of energy but little or no additional nutritional value. Because empty calories do not provide any other nutritional value, you have to eat other foods to get the necessary nutrients. This ultimately leads to overeating." (see link3)

The distribution of sugar_g is not so skewed and only has one peak. So, I will just use `sugar_g` not log10(sugar_g). The range is 83, and there are 473 variables. I calculated the binwidth should be around 3.8. The median is around 39. Also, as we can see the relationship between sugar and `calories_log10` in the scatterplot, it's positive and linear. It's a very strong relationship. The correlation between them is around 0.829, which is very close to 1. Regarding the confidence interval, the confidence interval level I chose is 95%. 0 is not in the range. There must be some connection between them. We can infer that a 95% bootstrap confidence interval for the slope of `sugar_g` can be calculated as the range [0.00772, 0.00866].



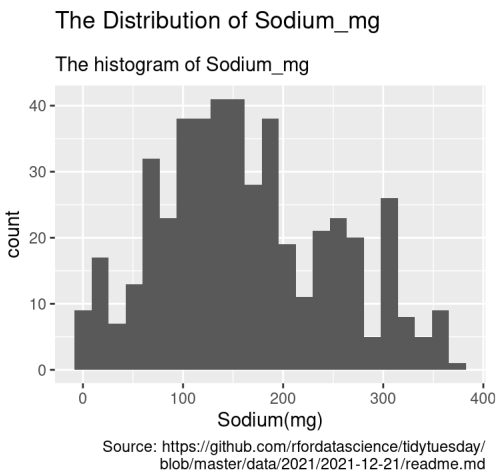


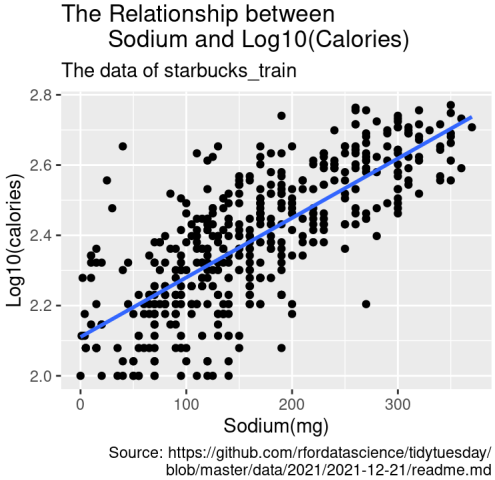

```
## # A tibble: 1 × 1
##   `cor(sugar_g, calories_log10)`
##                            <dbl>
## 1                          0.829
```

```
## # A tibble: 1 × 2
##     lower   upper
##     <dbl>   <dbl>
## 1 0.00772 0.00866
```

Then, why did I choose sodium? As the website creekside puts it, “several studies have shown that a high sodium diet can actually cause you to drink less water and be more hungry, which could then lead to overeating and more weight gain.”(see link4)

The distribution of `sodium_mg` is not skewed, symmetric, and only has one peak. So, I will just use `sodium_mg` not log10(sodium_mg). The range is 370, and there are 473 variables. I calculated the binwidth should be around 17. The median is around 150. Also, as we can see the relationship between sodium and `calories_log10` in the scatterplot, it’s positive and linear. It’s a relatively strong relationship. The correlation between them is around 0.755, which is quite close to 1. Regarding the confidence interval, the confidence interval level I chose is 95% too just as sugar. 0 is not in the range. There must be some connection between sodium and calories. We can infer that a 95% bootstrap confidence interval for the slope of `sodium_mg` can be calculated as the range [0.00157, 0.00182].

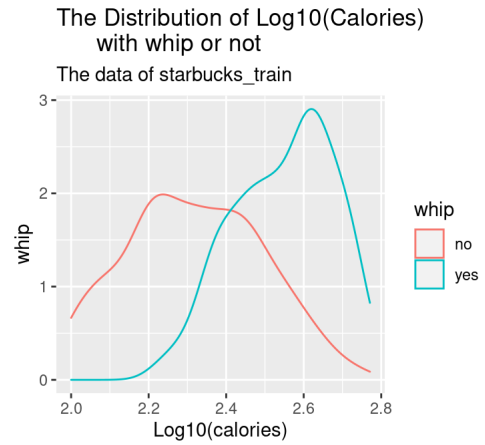


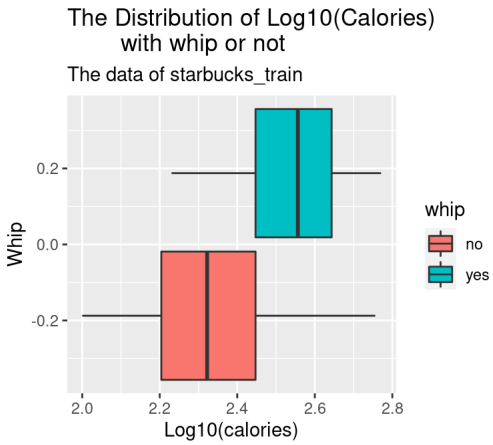



```
## # A tibble: 1 × 1
##   `cor(sodium_mg, calories_log10)`
##                              <dbl>
## 1                            0.755
```

```
## # A tibble: 1 × 2
##     lower   upper
##     <dbl>   <dbl>
## 1 0.00157 0.00182
```

For whip, as we can see from the density and boxplot, whip has a really strong impact on calories. The median `calories_log10` for coffee with whip is around 2.32. Most `calories_log10` are below 2.4. However, the median `calories_log10` for coffee with no whip is around 2.56. Most `calories_log10` are above 2.4. There is a big difference on log10(calories) for drinks with whip or not.As a result, `whip` is a really good explanatory variable.





All in all, these are some of my background and scientific reasons for why I choose `whip`, `sodium_mg`, and `sugar_g`. The next step I do is to create a model for estimating `clories_log10`. The predicted model is:

$$\hat{calories\_log10} = 2.03598 + 0.00529 * sugar\_g + 0.00058 * sodium\_mg + 0.12485 * whipyes$$

I also do some interpretation for each slope to better understand the model. Holding the other variables in the model the same, when sugar increases by 1g, the calories is expected to increase by a factor of 1.012Kcal, on average. Holding the other variables in the model the same, when sodium increases by 1mg, the calories is expected to increase by a factor of 1.001Kcal on average. Holding the other variables in the model the same, the calories of beverages of whip is expected to be 0.12485Kcal higher, on average.

```
## # A tibble: 4 × 5
##   term        estimate std.error statistic  p.value
##   <chr>          <dbl>     <dbl>     <dbl>    <dbl>
## 1 (Intercept) 2.04     0.00946      215.   0
## 2 sugar_g     0.00529  0.000319      16.6  4.58e-49
## 3 sodium_mg   0.000583 0.0000717      8.13 3.94e-15
## 4 whipyes     0.125    0.00919       13.6  1.12e-35
```

```
## # A tibble: 1 × 1
##   adj.r.squared
##           <dbl>
## 1         0.800
```

To evaluate whether this model is good or not, I did some calculations. The adjusting R_square is around 0.8. We can know that 80% of variability in the response variable can be explained by the regression model. The root mean square error is around 0.0871. 0.806(range) is 9.25 times more than 0.0871, which is really great.

By using statistical methods and evaluating the calculation results, the model is pretty good overall.

```
## # A tibble: 1 × 3
##   .metric .estimator .estimate
##   <chr>   <chr>          <dbl>
## 1 rmse    standard      0.0871
```

```
## # A tibble: 1 × 2
##     min   max
##   <dbl> <dbl>
## 1     2  2.81
```

Below, we made a table to display the r-squared and RMSE in our models.

Model	Training R Square adjusted	Testing RMSE
Liuyixin Shao	0.9306172	0.05284181
Cole Smidt	0.915	0.0628
Sophia Lan	0.8000315	0.08713573

Results

With the results of the r-squared and RMSE of each of our models, we have decided that the model made by Liuyixin best models the data. It has the highest adjusted r-squared and the lowest testing RMSE.

Using this model, the fitted model equation we came up with for `calories_log10` for the whole dataset is:

$$\hat{calories\_log10} = 1.99 - 0.0342 * sizeshort - 0.0121 * sizetall - 0.0397 * sizetrenta - 0.0325 * sizeventi(cold)$$
$$+0.0181 * sizeventi(hot) + 0.0068 * total\_carbs\_g + 0.0141 * total\_fat\_g$$

```
## # A tibble: 8 × 3
##   term            estimate undo_log
##   <chr>              <dbl>    <dbl>
## 1 (Intercept)      1.99      98.3
## 2 sizeshort       -0.0342     0.924
## 3 sizetall        -0.0121     0.972
## 4 sizetrenta      -0.0397     0.913
## 5 sizeventi(cold) -0.0325     0.928
## 6 sizeventi(hot)   0.0181     1.04
## 7 total_carbs_g    0.00682    1.02
## 8 total_fat_g      0.0141     1.03
```

Converting `calories` into a logarithmic scale helps us to have a normal distribution of our response variable, which enables us to build a better model. However, it is the variable `calories` that people truly care about when analyzing the contents of a drink. Therefore, we also applied our fitted model for `calories`, which displays the true number of calories in the dataset. The equation we get is:

$\hat{calories} = 27.11 - 7.57 * sizeshort - 7.43 * sizetall - 19.30 * sizetrenta - 0.88 * sizeventi(cold) + 18.73 * sizeventi(hot)$
$+3.86 * total\_carbs\_g + 9.69 * total\_fat\_g$

```
## # A tibble: 8 × 2
##   term            estimate
##   <chr>              <dbl>
## 1 (Intercept)       27.1
## 2 sizeshort         -7.57
## 3 sizetall          -7.43
## 4 sizetrenta       -19.3
## 5 sizeventi(cold)   -0.876
## 6 sizeventi(hot)    18.7
## 7 total_carbs_g      3.86
## 8 total_fat_g        9.69
```

```
## adj.r.squared for calories_log     adj.r.squared for calories
##                      0.9278275                      0.9787638
```

This means, holding the other variables in the model the same, when the total carbohydrates in the beverage increases by 1 gram, we can expect an increase of 3.86 kCal in total calories in the beverage, on average. When the total fat increases by 1 gram, we can expect an increase of 9.69 kCal in total calories in the beverage, on average. Both of these values are extremely close to the actual number of calories contained in a gram of carbohydrates, 4 kCal, and fat, 9 kCal. Holding the other variables in the model the same, if a customer orders a drink with “short” size, the calories inside will be expected to decrease 7.57 kCal, on average, while “tall” size to decrease by 7.43 kCal, “trenta” size to decrease by 19.30 kCal, “venti(cold)” size to decrease by 0.88 kCal, and “venti(hot)” size to increase by 18.73 kCal, on average. When total carbohydrates and total fat contained in the beverage is 0, we expect the total calories in a “short” size Starbucks beverage to be 19.54 kCal, while in a “tall” size Starbucks beverage to be 19.68 kCal, in a “trenta” size Starbucks beverage to be 7.81kCal, in a “venti(cold)” size Starbucks beverage to be 26.23 kCal, and in a “venti(hot)” size Starbucks beverage to be 45.84 kCal, on average.

Our regression model, when compared to the actual value of calories, has an adjusted r-squared value of 0.979, meaning it explains 97.9% of variability of the `calories`, which is even better than the r-squared modeling `calories_log`, where it explains 92.8% of the variability. This is a very accurate model of the data. The variables chosen in this model, `size`, `total_fat_g`, and `total_carbs_g`, were able to model this data the best due to their close relationship with the number of calories in a drink. Size is extremely important in determining the total calories, as the larger a drink is, the more calories it will have. Calories in “trenta” size beverages would perform slightly anomalously because this size is only served for a small number of beverages, meaning that their distribution is skewed when compared to the three other main sizes. Other than this, a large drink on average will always have more calories than a small drink of the same type. For example, in our dataset the vanilla sweet cream cold brew with no milk has 100 calories in the tall size, 110 in the grande, 200 in the venti, and 210 in the trenta. When it comes to the other two variables, both fat and carbohydrates have a strong linear relationship with calories. This relationship means that modeling the total number of calories in a given drink by these two numerical variables results in a well-performing model. Also, the correlation between fat and carbohydrate themselves is weak, proving that picking these two variables won’t cause collinearity to our model. Therefore, it’s the close relationships between the explanatory and response variables as well as the weak correlation between the explanatory variables themselves that builds this extremely accurate model, as can be seen by the r-squared value and the RMSE of the testing set.

What we have found with this data is that the calories in a Starbucks drink vary widely based on a number of factors. We have found the best way to measure the amount of calories in a drink is by using the amount of fat, the amount of carbohydrates, and the size of the drink. This allows customers like yourself of Starbucks to get a better understanding of what is contained within the drink they are consuming.

Bibliography:

Link1:https://github.com/rfordatascience/tidytuesday/blob/master/data/2021/2021-12-21/readme.md (https://github.com/rfordatascience/tidytuesday/blob/master/data/2021/2021-12-21/readme.md) & https://globalassets.starbucks.com/assets/91939428ecd94155a58b8765b6397c87.pdf (https://globalassets.starbucks.com/assets/91939428ecd94155a58b8765b6397c87.pdf)

Link2:https://www.rd.com/article/starbucks-coffee-sizes-explained/ (https://www.rd.com/article/starbucks-coffee-sizes-explained/)

Link3:https://www.creeksidefamilypractice.com/blog/salty-foods-how-sodium-affects-your-weight\ (https://www.creeksidefamilypractice.com/blog/salty-foods-how-sodium-affects-your-weight\)

Link4:https://drjessicablack.com/relationship-sugar-calories-health/ (https://drjessicablack.com/relationship-sugar-calories-health/)