Lecture 2.1 - Association activity

Student

2025-11-04

Planning

Expectations

You should develop some expectations based on some pre-existing ideas about how you think the data is related. Pick `Calories` and another quantitative variable and write down how you expect the two to be related.

Table 1: Outliers

Category	Item	Total.Fat	Calories
Chicken & Fish	Chicken McNuggets (40 piece)	118	1880

> Write down your expectation and provide a brief reason for the expectation. Remember to describe your expectation about the relationship between the two variables in terms of direction, form, strength, and outliers.

Let's pick `Total.Fat`. Probably the relationship is positive, somewhat linear, with a moderate strength. Outliers may include very fatty items or ones that are high in sugar but low in fat.

Direction of the relationship

Make a decision about which variable is the outcome variable you care about and which is the variable that predicts it. Write down your choice.

Outcome variable is `Calories`, predictor variable is `Total.Fat`. Fat is one component of energy a food may contain (a calorie).

Investigation

Distribution displays

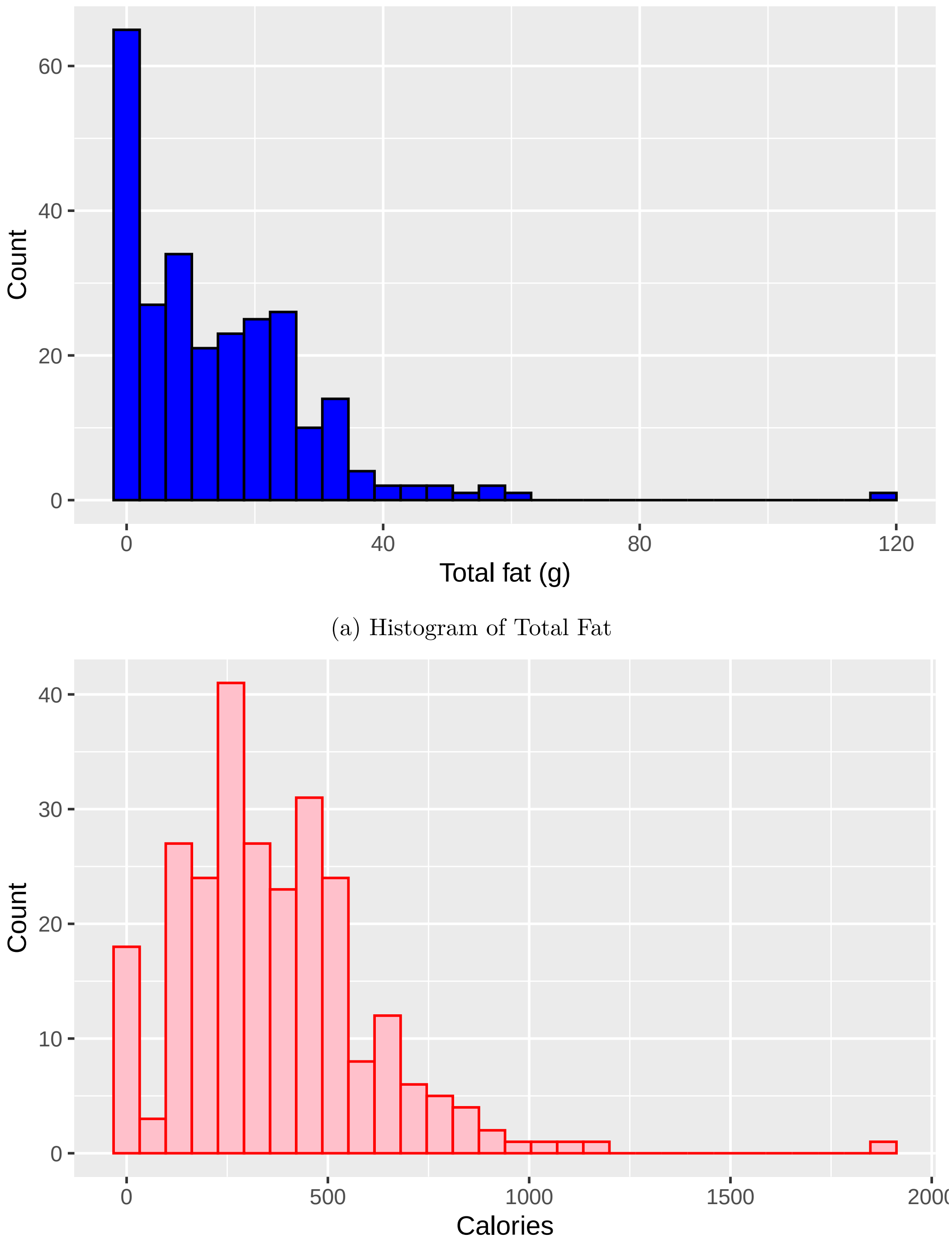
Histograms

- Make histograms of your two variables using `ggplot()` and `geom_histogram()`. Do you see anything unusual? What do you think these distributions indicate about the possible relationship between the two?

It seems that both are right skewed. It is a bit hard to tell what the relationship will be from the histograms. As identified in the previous lecture, the outlier here is 40 piece chicken nuggets.

Scatterplot

- Make a high-quality scatterplot of the two variables using `geom_point()`. You can add a smoother to the scatterplot by adding a `geom_smooth()` layer to your `ggplot`.



(a) Histogram of Total Fat

(b) Histogram of Calories

Figure 1: Histogram of predictor and response variables

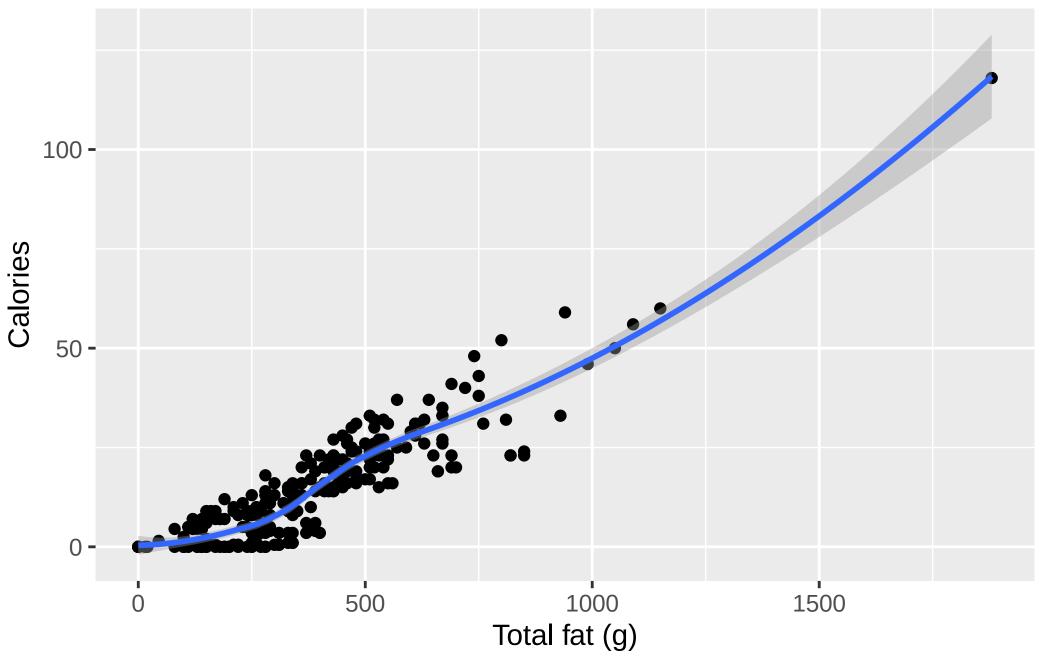


Figure 2: Relationship between Calories and Total Fat

Describe the relationship between the two variables using the terms we learned in class.

- *Direction: positive*
- *Form: mostly linear*
- *Strength: relatively strong, but with some possible group issues*
- *Outliers: the already noted outlier, chicken nuggets (40 piece)*

  Does the relationship match your expectations? Why or why not? Write some notes about comparing expectations vs. reality.

Yes, this fairly closely matches what we expected. Generally speaking, the more fat, the more calories. This suggests most of the calories come from the fat in McDonald's menu items.

Correlation

Find correlation result – does it match the scatterplot? Your expectations? Why or why not?

The RStudio code for correlations is:

```
cor(x, y, use="complete.obs")
cor(mcdonalds$Calories, mcdonalds$Total.Fat)
```

```
[1] 0.9044092
```

- Note: you will need to replace `x` and `y` in the above line of code with the variables of interest. Remember to directly specify a variable, it should be in the `<dataset name>$<variabe>` format.

Yes, it does match the expectations. It does look highly linear.

Analysis

Outliers

If you have any outliers, identify them. Do you think they should be excluded from your analysis? Why or why not? If you remove the outliers (via the `filter()` verb), does it change your correlation? The shape of the smoother?

As identified above, is is the chicken nugget menu item. I think it should be excluded, it is not the kind of item a regular person usually orders - it is for sharing.

```
mcdonalds.nooutlier <- mcdonalds %>%
  filter(Calories < 1500)

cor(mcdonalds.nooutlier$Calories, mcdonalds.nooutlier$Total.Fat)
```

```
[1] 0.8863115
```

The correlation does not change very much, it is already on a line of best fit so this is not surprising. The item type is just a lot more of the same kind of food in "normal" menu items.

Here we can see that the relationship is still strong without the outlier. However, this plot also draws our focus to the cases with 0 fat that sometimes have a lot of calories that we did not notice earlier.

From here, we need to consider if these are really part of the relationship we are interested in.

Table 2: No fat but high calorie outliers

Category	Item	Total.Fat	Calories
Snacks & Sides	Side Salad	0	20
Snacks & Sides	Apple Slices	0	15
Beverages	Coca-Cola Classic (Small)	0	140
Beverages	Coca-Cola Classic (Medium)	0	200
Beverages	Coca-Cola Classic (Large)	0	280
Beverages	Coca-Cola Classic (Child)	0	100
Beverages	Diet Coke (Small)	0	0
Beverages	Diet Coke (Medium)	0	0
Beverages	Diet Coke (Large)	0	0
Beverages	Diet Coke (Child)	0	0
Beverages	Dr Pepper (Small)	0	140
Beverages	Dr Pepper (Medium)	0	190
Beverages	Dr Pepper (Large)	0	270
Beverages	Dr Pepper (Child)	0	100
Beverages	Diet Dr Pepper (Small)	0	0
Beverages	Diet Dr Pepper (Medium)	0	0
Beverages	Diet Dr Pepper (Large)	0	0
Beverages	Diet Dr Pepper (Child)	0	0
Beverages	Sprite (Small)	0	140
Beverages	Sprite (Medium)	0	200
Beverages	Sprite (Large)	0	280
Beverages	Sprite (Child)	0	100
Beverages	Fat Free Chocolate Milk Jug	0	130
Beverages	Minute Maid 100% Apple Juice Box	0	80
Beverages	Minute Maid Orange Juice (Small)	0	150
Beverages	Minute Maid Orange Juice (Medium)	0	190
Beverages	Minute Maid Orange Juice (Large)	0	280
Beverages	Dasani Water Bottle	0	0
Coffee & Tea	Iced Tea (Small)	0	0
Coffee & Tea	Iced Tea (Medium)	0	0
Coffee & Tea	Iced Tea (Large)	0	0
Coffee & Tea	Iced Tea (Child)	0	0
Coffee & Tea	Sweet Tea (Small)	0	150
Coffee & Tea	Sweet Tea (Medium)	0	180
Coffee & Tea	Sweet Tea (Large)	0	220
Coffee & Tea	Sweet Tea (Child)	0	110
Coffee & Tea	Coffee (Small)	0	0
Coffee & Tea	Coffee (Medium)	0	0
Coffee & Tea	Coffee (Large)	0	0
Coffee & Tea	Nonfat Latte (Small)	0	100
Coffee & Tea	Nonfat Latte (Medium)	0	130
Coffee & Tea	Nonfat Caramel Latte (Small)	0	200
Coffee & Tea	Nonfat Caramel Latte (Medium)	0	250
Coffee & Tea	Nonfat Hazelnut Latte (Small)	0	200
Coffee & Tea	Nonfat Hazelnut Latte (Medium)	0	250
Coffee & Tea	Nonfat French Vanilla Latte (Small)	0	190
Coffee & Tea	Nonfat French Vanilla Latte (Medium)	0	240
Coffee & Tea	Nonfat Latte with Sugar Free French Vanilla Syrup (Small)	0	140
Coffee & Tea	Nonfat Latte with Sugar Free French Vanilla Syrup (Medium)	0	170

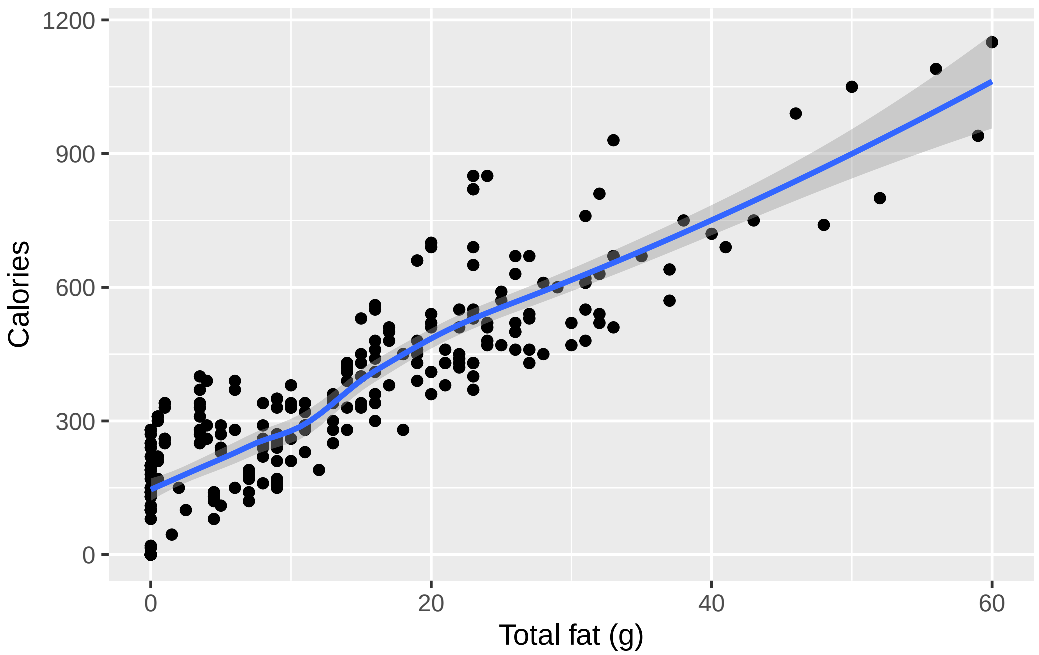


Figure 3: Relationship between Calories and Total Fat

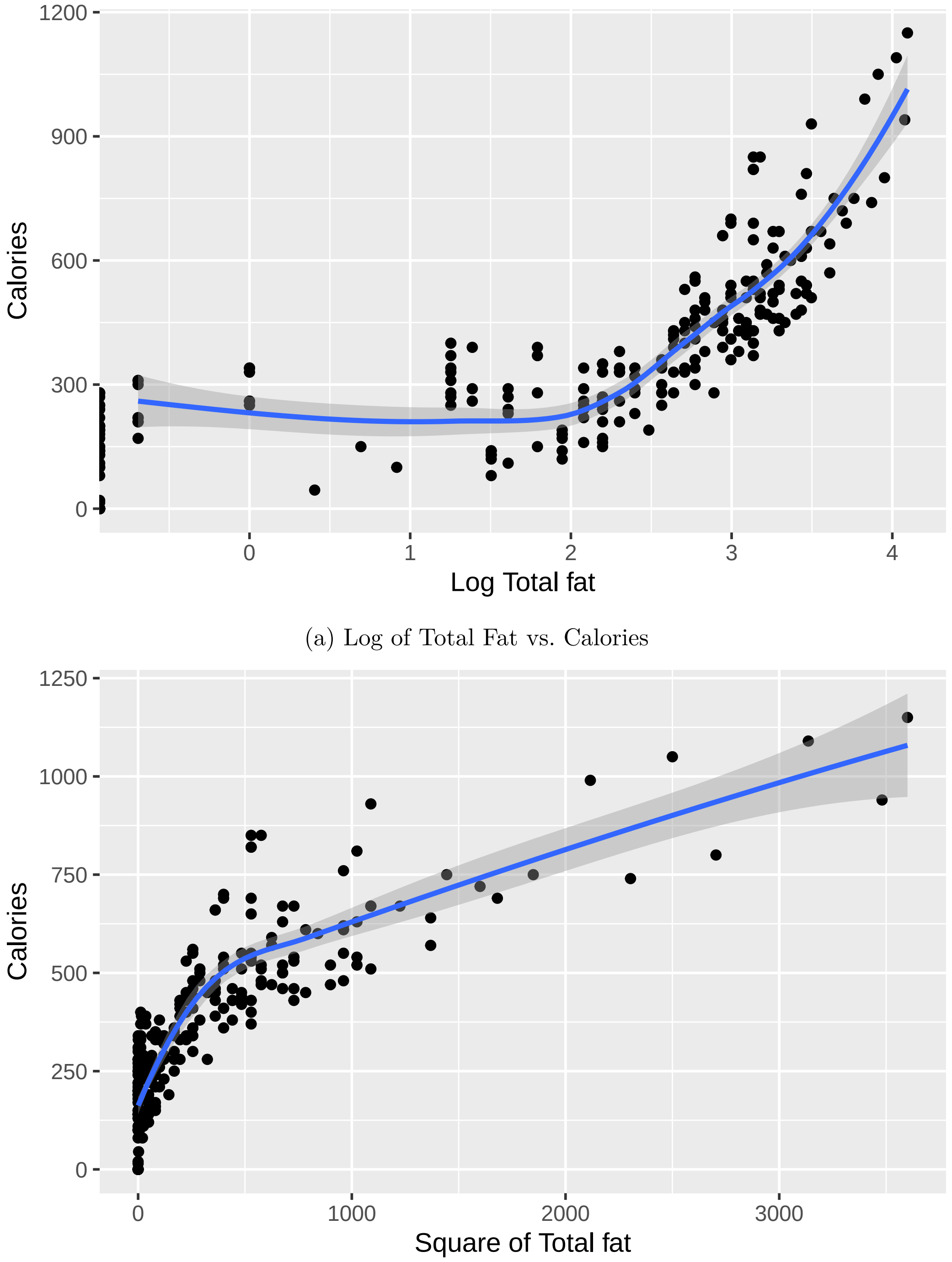
Transformations

- Could the relationship of your two chosen variables benefit from being re-expressed? Try a few re-expressions and see how it affects the relationship and correlation.
  - Hint: remember, you can use the `mutate()` verb in a piped command, such as: `mutate(logcals = log(Calories))`

Since the relationship was already linear, a reexpression will probably not help here. A view of the reexpressed graphs confirms this.

Conclusion

Overall, summarize what you have learned about the relationship of your two variables.



(a) Log of Total Fat vs. Calories

(b) Square of Total Fat vs. Calories

Figure 4: Log and squared reexpressions