AE 12: Multiple linear regression cont'd

Your Name

2021-09-29

```
library(tidyverse)
library(broom)
library(patchwork)
```

For this analysis, we will use the LEGO data set you analyzed in Exam 01.

The data for this analysis includes information about LEGO sets from themes produced January 1, 2018 and September 11, 2020. The data were originally scraped from Brickset.com, an online LEGO set guide.

You will work with data on about 400 randomly selected LEGO sets produced during this time period. The primary variables are interest in this analysis are

- `Pieces`: Number of pieces in the set from brickset.com.
- `Amazon_Price`: Amazon price of the set scraped from brickset.com (in U.S. dollars)
- `Size`: General size of the interlocking bricks (Large = LEGO Duplo sets - which include large brick pieces safe for children ages 1 to 5, Small = LEGO sets which- include the traditional smaller brick pieces created for age groups 5 and - older, e.g., City, Friends)

The goal of this analysis is to predict the Amazon price based on the number and size of the pieces. We will only include observations that have recorded values for all three variables.

```
legos <- read_csv("data/lego-sample.csv") %>%
  filter(!is.na(Size), !is.na(Pieces), !is.na(Theme))
```

Amazon Price vs. Pieces and Size

We'll start with a linear regression model using the number and size of piece and their interaction to predict the Amazon.com price and output the results.

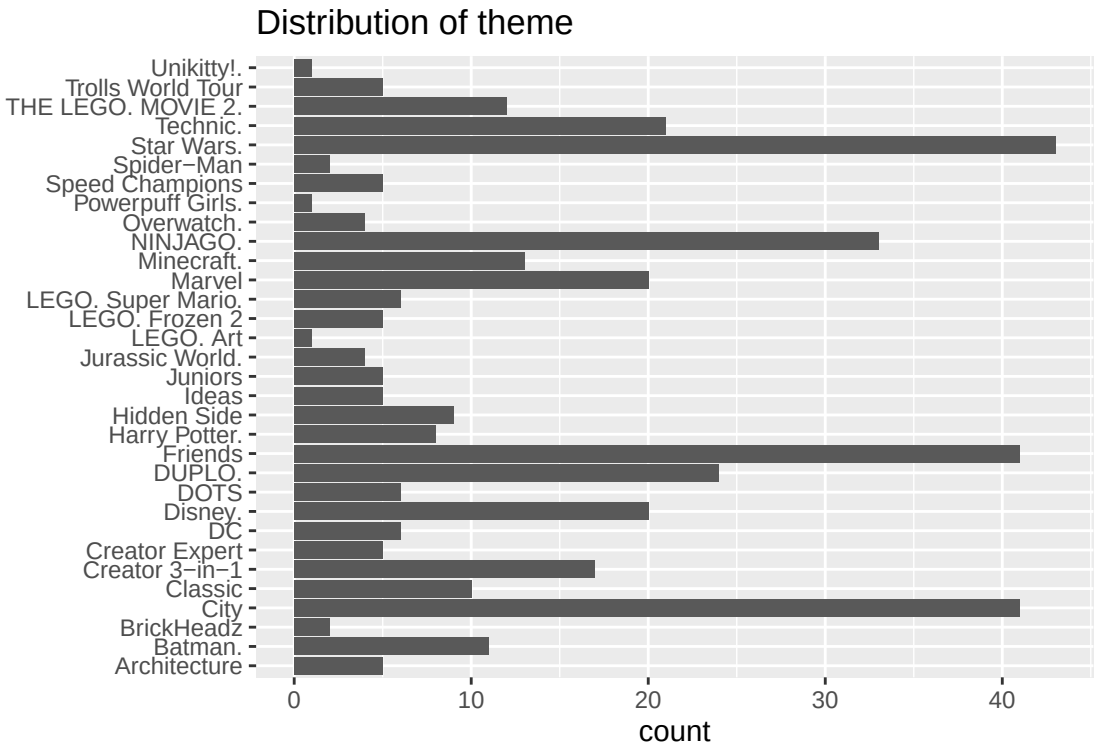
```
## fit linear model
```

1. The coefficient for `SizeSmall` is an adjustment on the [FILL IN] for LEGO sets with small pieces?
2. The coefficient for the interaction term is an adjustment on the [FILL IN] for LEGO sets with small pieces?
3. Write the regression equation for sets with large pieces.
4. Write the regression equation for sets with small pieces.

Amazon Price vs. Pieces, Size, and Theme

Before fitting the new model, let's take a look at the distribution Theme

```
ggplot(data = legos, aes(x = Theme)) +
  geom_bar() +
  labs(x = "",
       title = "Distribution of theme") +
  coord_flip()
```



There are a lot of categories of Theme, and many take very few values. Rarely do we need this many levels of a categorical variable in a model. Having this many levels can use a lot of degrees of freedom while not providing much useful information in the interpretation.

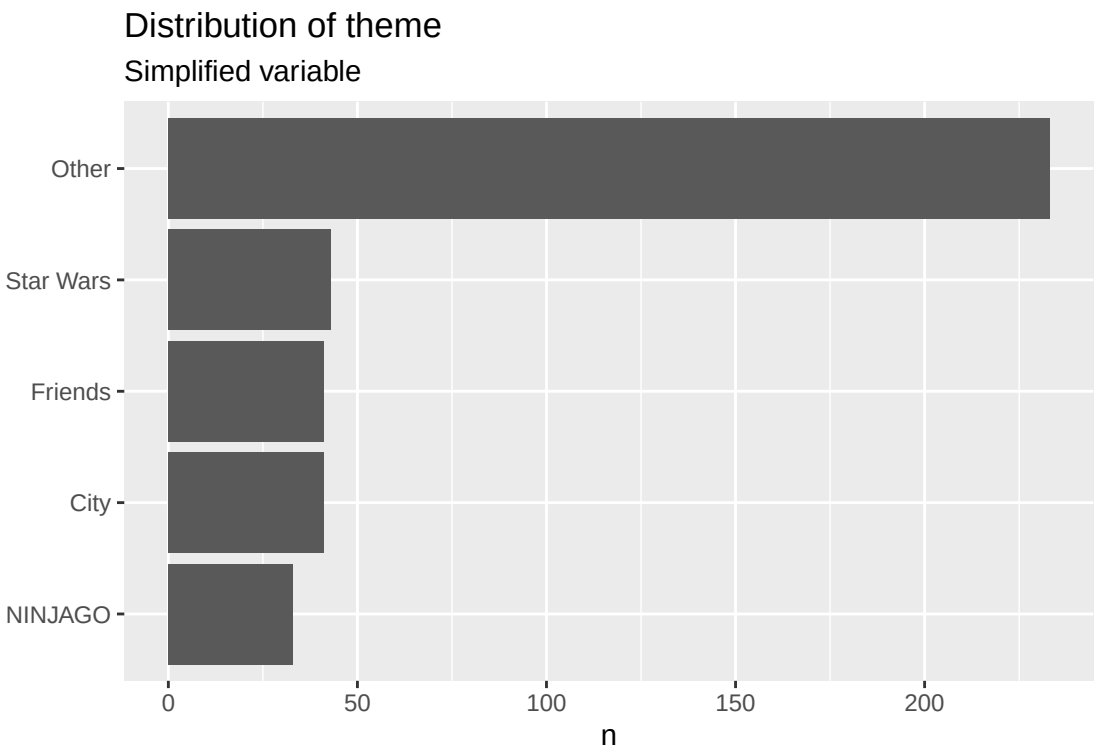
Given this, we will create a simplified version of this variable that only includes 5 categories (could be a different value). There are multiple ways to simplify the variable, but we'll use `fct_lump` which simplifies based on the number of observations.

```
# simplify categories
legos <- legos %>%
  mutate(theme_new = fct_lump(Theme, 4, other_level = "Other",
                              ties.method = "first")) #how to break ties

# remove extra characters (from trademark symbols)
legos <- legos %>%
  mutate(theme_new = str_replace(theme_new, "\UFFFD" ,""))
```

Now let's look at `theme_new`, the simplified version of theme.

```
legos %>%
  count(theme_new, sort = T) %>%
ggplot(aes(x = fct_reorder(theme_new, n), y = n)) +
  geom_col(stat = "identity") +
  labs(x = "",
       title = "Distribution of theme",
       subtitle = "Simplified variable") +
  coord_flip()
```



Fit the model using pieces, size, and `theme_new`.

1. What is the baseline level of `theme_new`?
2. What is the interpretation of the coefficient for Star Wars?
3. What is the difference in the predicted Amazon.com price between a Friends set with 500 pieces and a Star Wars set with 100 pieces, each with small pieces?

Mean-centered variables

1. Create a new variable for the mean-centered value for the number of pieces. Then refit the model from the previous exercise using the mean-centered value of pieces instead of the original variable.

```
## add code
```

2. Interpret the slope of pieces in the context of the data.
3. Which LEGO sets are represented by the intercept?