Lab 01: Ikea furniture

Simple linear regression

Dav King

2022-09-09

Setup

Load packages and data:

```
library(tidyverse)
library(tidymodels)
library(ggridges)
library(cowplot)
ikea <- read_csv("data/ikea.csv")
```

Exercises

Exercise 1

```
glimpse(ikea)
```

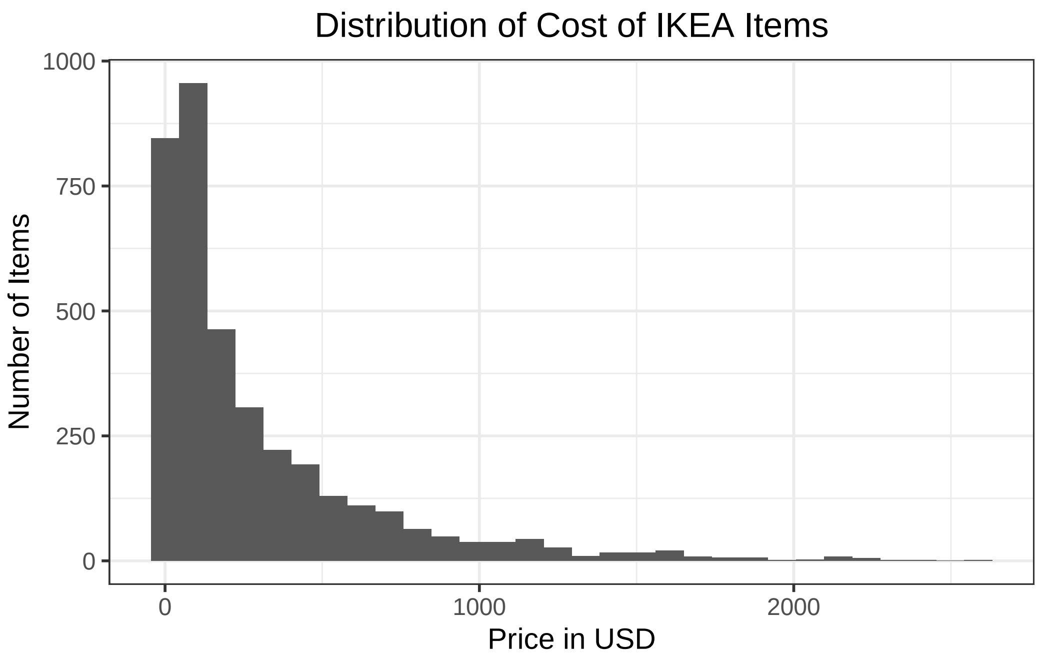
```
Rows: 3,694
Columns: 13
$ X1                <dbl> 0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15~
$ item_id           <dbl> 90420332, 368814, 9333523, 80155205, 30180504, 10122~
$ name              <chr> "FREKVENS", "NORDVIKEN", "NORDVIKEN / NORDVIKEN", "S~
$ category          <chr> "Bar furniture", "Bar furniture", "Bar furniture", "~
$ sellable_online   <lgl> TRUE, FALSE, FALSE, TRUE, TRUE, TRUE, TRUE, TRUE, TR~
$ link              <chr> "https://www.ikea.com/sa/en/p/frekvens-bar-table-in-~
$ other_colors      <chr> "No", "No", "No", "Yes", "No", "No", "No", "No", "No~
$ short_description <chr> "Bar table, in/outdoor,          51x51 cm", "Bar tab~
$ designer          <chr> "Nicholai Wiig Hansen", "Francis Cayouette", "Franci~
$ depth             <dbl> NA, NA, NA, 50, 60, 45, 44, 50, 44, NA, 44, 45, 47, ~
$ height            <dbl> 99, 105, NA, 100, 43, 91, 95, NA, 95, NA, 103, 102, ~
$ width             <dbl> 51, 80, NA, 60, 74, 40, 50, 50, 50, NA, 52, 40, 46, ~
$ price_usd         <dbl> 71.55, 268.65, 565.65, 18.63, 60.75, 93.15, 34.83, 5~
```

The `ikea` dataset has 3694 observations and 13 variables.

Exercise 2

```r
ggplot(ikea, aes(x = price_usd)) +
  geom_histogram() +
  labs(x = "Price in USD", y = "Number of Items",
       title = "Distribution of Cost of IKEA Items") +
  theme_bw() +
  theme(plot.title = element_text(hjust = 0.5))
```

```
`stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



Exercise 3

The distribution of prices is highly right-skewed, with a modal peak around 100-150 and a spread from ~0 to 2500. There are plenty of outliers - really any value over 1000 probably is, but there are a number of items that cost over $2000 that would all easily be considered outliers. Because the data are skewed, the median is much more representative than the mean - the presence of highly weighted outliers would cause the mean to be much higher than where most of the data sit, whereas the median is unaffected by these outliers and would instead show the true “middle” value of the data.

Exercise 4

```
ikea_sub <- ikea %>%
  filter(category %in% c("Tables & desks", "Beds",
                         "Bookcases & shelving units", "Sofas & armchairs"))
```

There are 1796 observations of 13 variables in the new dataset.

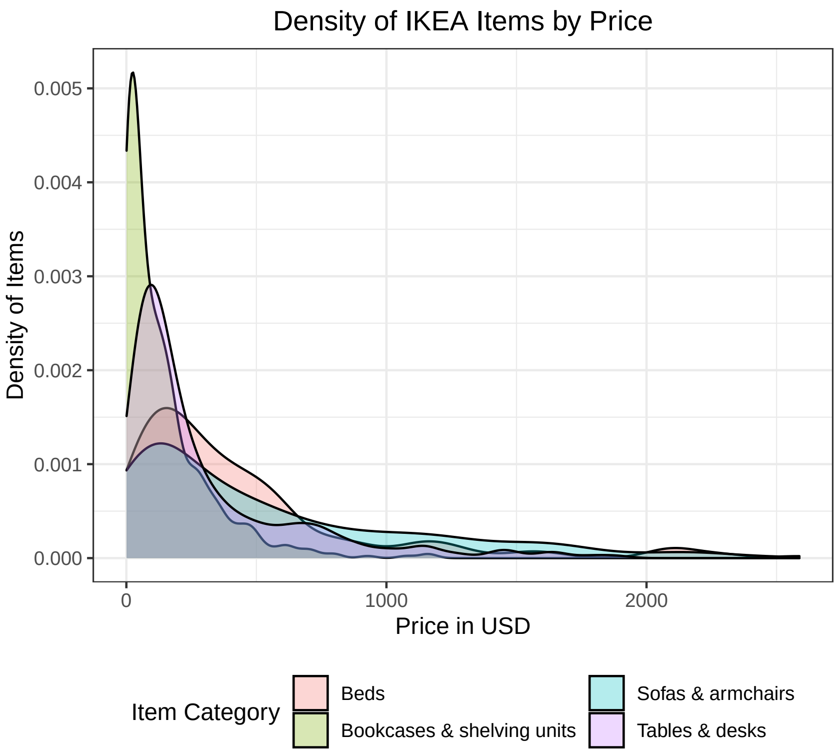
Exercise 5

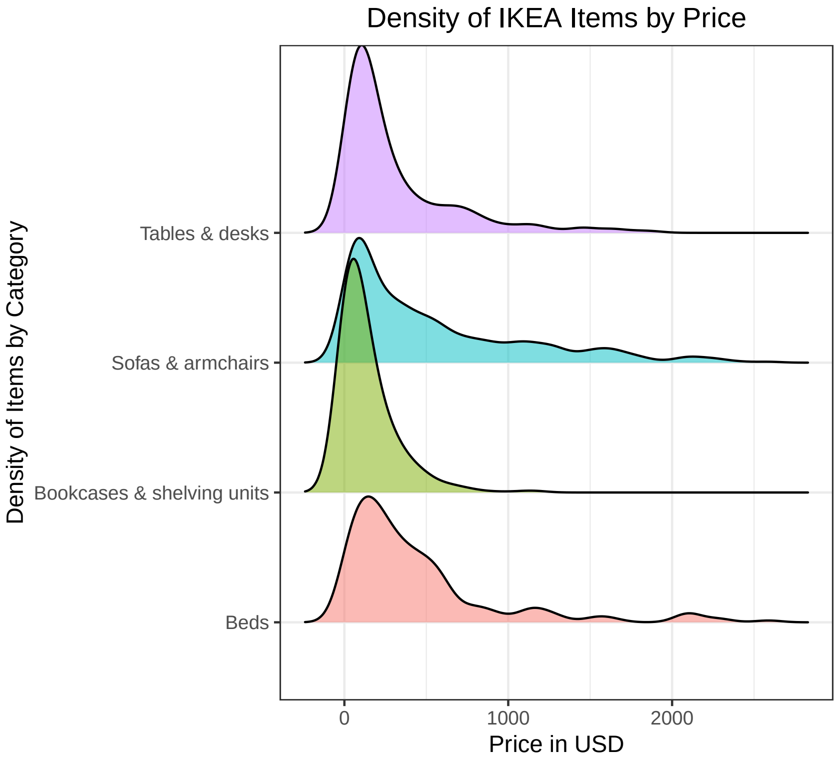
```
densityPlot <- ggplot(ikea_sub, aes(x = price_usd, fill = category)) +
  geom_density(alpha = 0.3) +
  theme_bw() +
  labs(x = "Price in USD", y = "Density of Items",
       title = "Density of IKEA Items by Price", fill = "Item Category") +
  theme(plot.title = element_text(hjust = 0.5), legend.position = "bottom") +
  guides(fill = guide_legend(nrow = 2))

densityRidgePlot <- ggplot(ikea_sub, aes(x = price_usd, y = category,
                                         fill = category)) +
  geom_density_ridges(alpha = 0.5) +
  theme_bw() +
  labs(x = "Price in USD", y = "Density of Items by Category",
       title = "Density of IKEA Items by Price") +
  theme(plot.title = element_text(hjust = 0.5), legend.position = "none")

plot_grid(densityPlot, densityRidgePlot, nrow = 2)
```

```
Picking joint bandwidth of 80.3
```


Density of IKEA Items by Price
Density of Items
0.005
0.004
0.003
0.002
0.001
0.000
0
1000
2000
Price in USD
Item Category
Beds
Bookcases & shelving units
Sofas & armchairs
Tables & desks

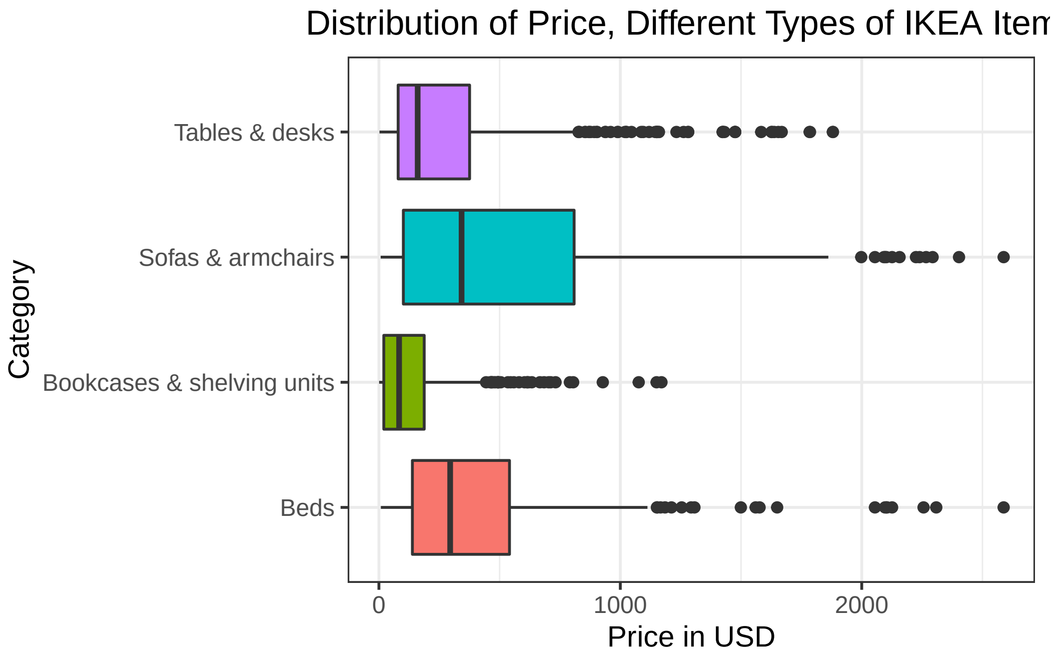

Density of IKEA Items by Price
Density of Items by Category
Tables & desks
Sofas & armchairs
Bookcases & shelving units
Beds
0
1000
2000
Price in USD

Exercise 6

We defined the fill within aesthetics because we wanted fill to take on different values based on a certain variable. We defined alpha within `geom` and not `aes`, however, because we wanted alpha to be a value we set ourselves and not a value dependent upon some characteristic of the data.

Exercise 7

```
ggplot(ikea_sub, aes(x = price_usd, y = category, fill = category)) +
  geom_boxplot() +
  labs(x = "Price in USD", y = "Category",
       title = "Distribution of Price, Different Types of IKEA Items") +
  theme_bw() +
  theme(legend.position = "none", plot.title = element_text(hjust = 0.5))
```



Exercise 8

My boxplot and the original density plot are both great at showing off where the data tend to be concentrated for each category, and especially their right-tailed skew and numerous outliers. The boxplot is better at showing off where the median lies for each category and explicitly designating which data points are outliers within each category. However, the density plot is better at showing off the sheer number of items concentrated at lower price points.

Exercise 9

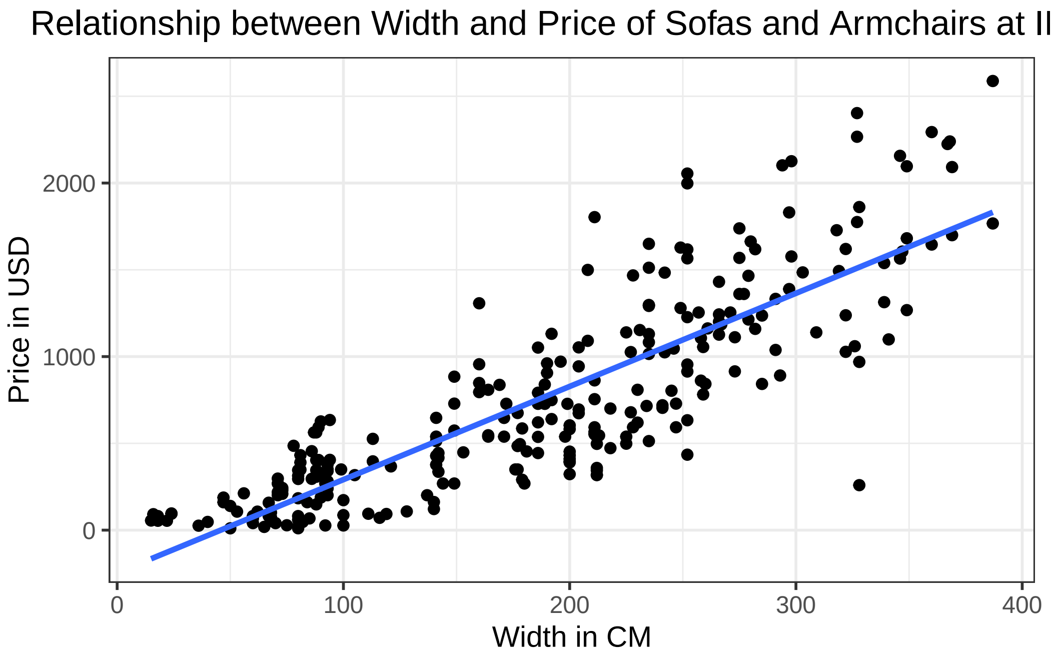
```r
sofas <- ikea_sub %>%
  filter(category == "Sofas & armchairs") %>%
  drop_na(width, price_usd)
```

There are 273 observations of 13 variables in `sofas`.

Exercise 10

```
ggplot(sofas, aes(x = width, y = price_usd)) +
  geom_point() +
  geom_smooth(method = "lm", se = F) +
  theme_bw() +
  labs(x = "Width in CM", y = "Price in USD",
title = "Relationship between Width and Price of Sofas and Armchairs at IKEA") +
  theme(plot.title = element_text(hjust = 0.5))
```

```
`geom_smooth()` using formula 'y ~ x'
```



There is a fairly strong positive linear relationship between width and price of sofas and armchairs at IKEA.

Exercise 11

```
priceModel <- linear_reg() %>%
  set_engine("lm") %>%
  fit(price_usd ~ width, data = sofas)
tidy(priceModel)
```

```
# A tibble: 2 x 5
  term        estimate std.error statistic  p.value
  <chr>          <dbl>     <dbl>     <dbl>    <dbl>
1 (Intercept)  -246.      43.0       -5.71 2.92e- 8
2 width           5.37     0.207     26.0  1.51e-75
```

Exercise 12

$$\widehat{price\_usd} = -245.671 + 5.368 \times width$$

Exercise 13

For each 1cm increase in the sofa or armchair's width, we would expect to see an increase in price of $5.368. The intercept is not meaningful - we would not expect a couch to be free, and we definitely can't have one with a width of -246 centimeters.

Exercise 14

```
glance(priceModel)$r.squared
```

```
[1] 0.7135975
```

```
aug <- augment(priceModel$fit)
rmse(aug, truth = price_usd, estimate = .fitted)$.estimate
```

```
[1] 311.694
```

The R^2 value for this model is 0.714, which means that 71.36% of the variance in price can be explained by the width of the couch/armchair - a very reasonable value. The RMSE of the model is 311.694, which means that $311.69 is the square root of the average of squared differences between the price that our model predicts based off of an item's width and its actual recorded price at that width. In other words, it is our average residual, or the absolute value by which our model errs in its predictions on average.

The R^2 value of this model suggests that the linear model we have created is a good fit for the data - our one predictor being able to explain over 70% of the variance in price suggests a good fit for the data. Unfortunately, the RMSE doesn't tell us much about whether this model is a good fit for the data - it would work a lot better if we had other models to compare it to, but a standalone RMSE can't tell us much. Still, from the good R^2 value, we can say that this model is a good fit for the data - not perfect, since we still cannot explain nearly 30% of the variance in price, but still a fairly strong predictor.

(Note: I did indeed know what the values were that I was writing about, I just wrote them with inline code instead to practice.)