Sneaky Machine Learning

Jayme Gerring, Brendan Ok, Pin-Yun Lin

5/9/2022

Abstract

blah blah blah

Introduction

Pulling in $70 billion in 2020, the sneaker market has a powerful influence within American retail. Because of the high demand for these sometimes rare and unique shoes, a powerful resale market has also emerged. The sneaker resale market was worth as much as $2 billion in 2019, a figure that has only increased as more and more players try to get in on the sometimes over 2000% profit margin earned from the rarest of sneakers.

As three certified 'sneakerheads' we were interested in using machine learning methods to accurately predict the premiums that result from reselling popular sneakers.

Price Premium is defined as:

$$Premium(\%) = \frac{ResalePrice(\$) - RetailPrice(\$)}{RetailPrice(\$)}$$

Why is this relevant? Premiums are a quick and simple benchmark to measure the profitability and desirability of a specific sneaker. Many characteristics, such as colorway[1], brand, size, and material can make or break a shoe sale. The physical characteristics of shoes are not the only determining factors for premiums, much like other retail goods, shoe sales have a seasonality component as well. This makes understanding the timing of a sale crucial. Premiums can demonstrate to resellers which characteristics make a shoe more profitable. Premiums can also be useful to buyers: based on characteristics, what price is a good deal and what prices border on irrational?

(Think about adding more)

Methodology

Part 1: Data Descriptions

The final dataset used in this project is located in `data/shoe_final.csv`

Scripts used to merge variables and clean data are located in `r/`

[1] Colorway is a term used to quickly sum up the colors of the sneakers, in our dataset we have colorway categorized as primary, secondary and tertiary colors.

The specific shoe data for this project was collected from the popular resale website StockX. The dataset contains the details of 99,956 orders of *Yeezy* and *Nike X Off-White* shoes made on StockX from September 2017 to February 2019. The variables associated with orders are: *Buyer Region* (State), *Order Date*, *Brand*, *Shoe Name*, *Retail Price*, *Sale Price*, *Release Date*, and *Size*.

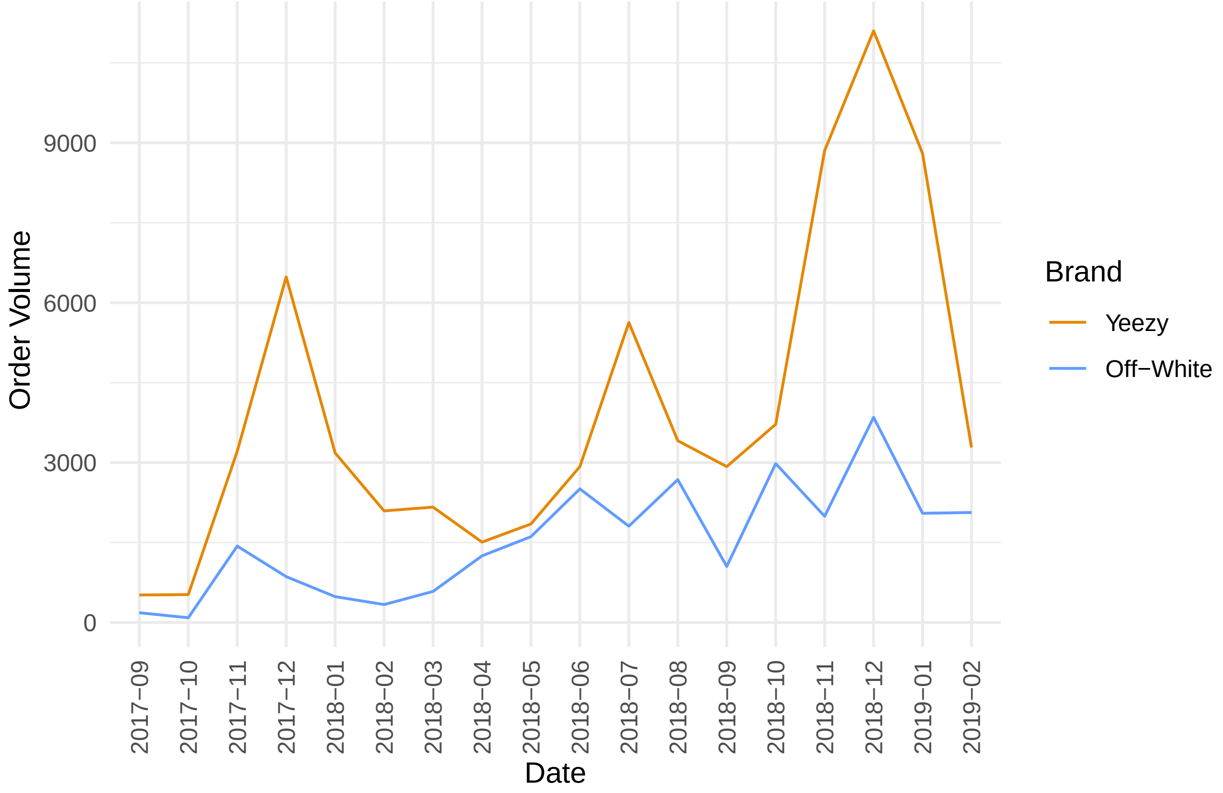
Premium was created from this initial dataset using the formula described in the previous section.

We collected additional variables regarding characteristics of each shoe including: *Material*, *Lace Type*. *Primary Color*, *Secondary Color*, and *Tertiary Color*,

Because certain buying choices could be reflective of economic conditions, we added the variables: *USA Monthly Retail Sales* (Monthly), *State Disposable Income per Capita* (Yearly), and *State Population* (Yearly)

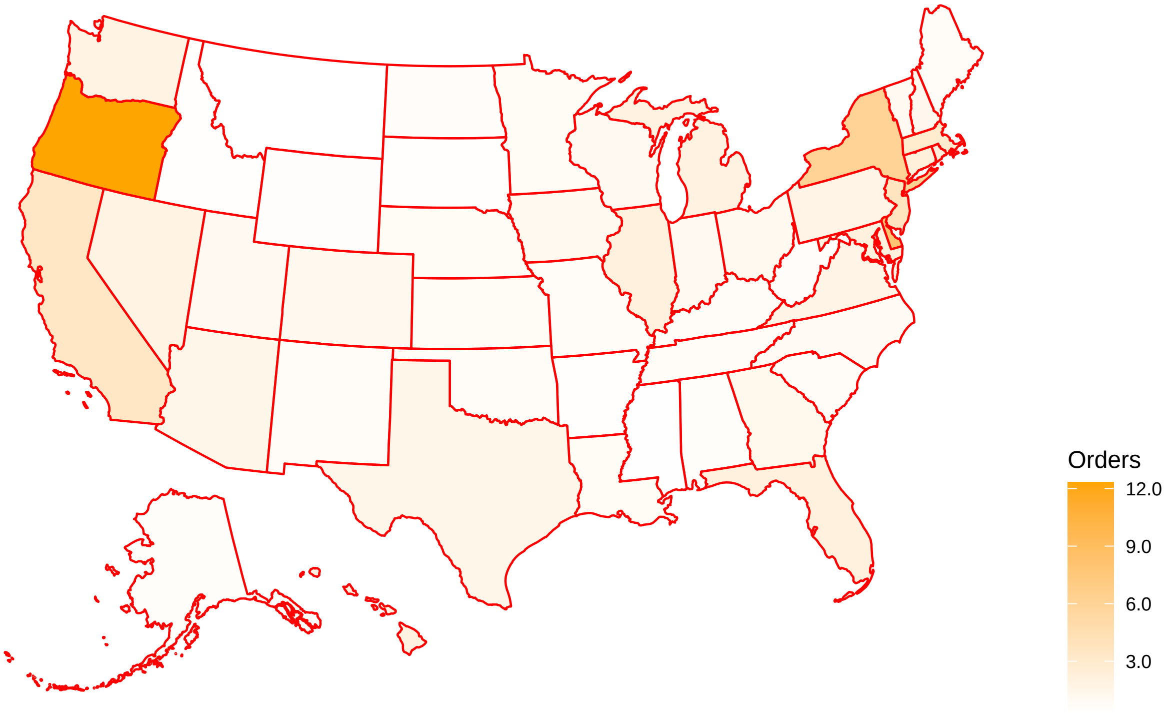
A quick glance at monthly order volume by brand (*Figure 1*), reveals a definite seasonal pattern, with orders spiking for both brands around the holiday season in both 2017 and 2018. The data also exhibit non-seasonal spikes in order numbers that appear to be linked to specific product release dates and restocks. For example, we believe the July 2018 spike in *Yeezy* orders could be associated with the late June release of the *350 V2 "Butter"*. It should be noted that the steep decline in orders around February 2019 is due to the data ending in the middle of the month.

Figure 1: Monthly Order Volume, Over Time



We also wanted to address any geographical component to order volume. After mapping orders, we noticed that Oregon had a disproportional share of orders not explained by population. As shown in *Figure 2*, when controlled for population, Oregon still seems to order the most sneakers out of any state. *Figure 2* displays orders in 2018, but the effect is still pronounced in 2017 and 2019. The maps for these two years have been included in the appendix as *Figure A* and *Figure B*, respectively.

Figure 2: 2018 Total Order Count per 10000 Persons



After looking at order volume, we then decided to turn our attention to premiums. Over the entire data set, *Nike X Off-White* has around a 284% premium and *Yeezy* has a premium of around 64%.

Plotting the average premium over time, we can see that there again appears to be a seasonality effect. *Figure 3* displays the average premium by brand over time. Interestingly, the average premium seems to dip for each brand around the holiday season. This effect could be due to a saturation of sellers trying to take advantage of the holiday season and new releases/restocks of shoes. The downward trend of premiums over time could be due to a variety of factors: possibly more people are selling on StockX over time, driving premiums down as sellers compete for consumers. Another factor driving down premiums could be that *Yeezy* and *Nike X Off-White* are putting out more stock to keep up with demand, driving premiums down on the demand side.

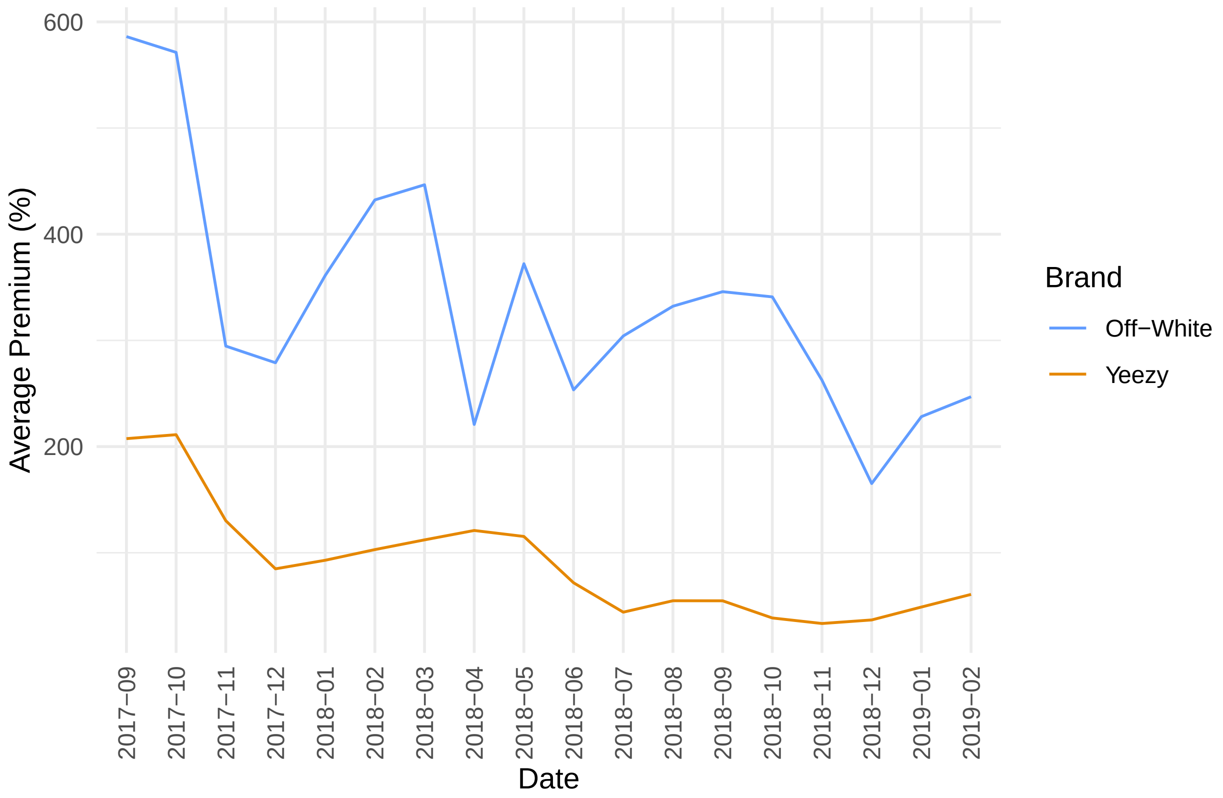


Figure 3: Monthly Average Premium, Over Time

Because of the geographical effects seen in Oregon with order volume in *Figure 2*. We decided to investigate the geographical effects of resale premiums. We found that in 2017, Kentucky had an unusually high average premium. In 2019, both Utah and Hawaii carried larger average premiums. Maps displaying the average premiums by state can be found in the appendix as *Figure C*, *Figure D*, and *Figure E*. We determined that the high average premium in Kentucky in 2017 was caused by a single sale of sneakers that carried a 2000% premium. The reasons for the higher average premiums in Utah and Hawaii appear to related to tastes. Both of these states had relatively small order volumes, and the majority of the sneakers purchased were *Nike X Off-White* which typically have higher premiums than *Yeezy*.

(We should drop variables from the CSV if we're not going to use them)