Sneaky Machine Learning

Jayme Gerring, Brendan Ok, Pin-Yun Lin

5/9/2022

Abstract

This report focuses on the question of premiums gained from reselling sneakers on the popular website StockX. We were interested in finding what characteristics determine premiums, that is the percent change in sale price on StockX compared to the shoe's original retail price. Using Random Forest Classification methods, we were able to determine that sale date and shoe size were the most important determining characteristics in maximizing premium. We also developed the secondary question examining the overrepresentation of Oregon in StockX orders, we determined that Portland's sneaker culture may be responsible for its outlier status in the data.

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?

[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.

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/`

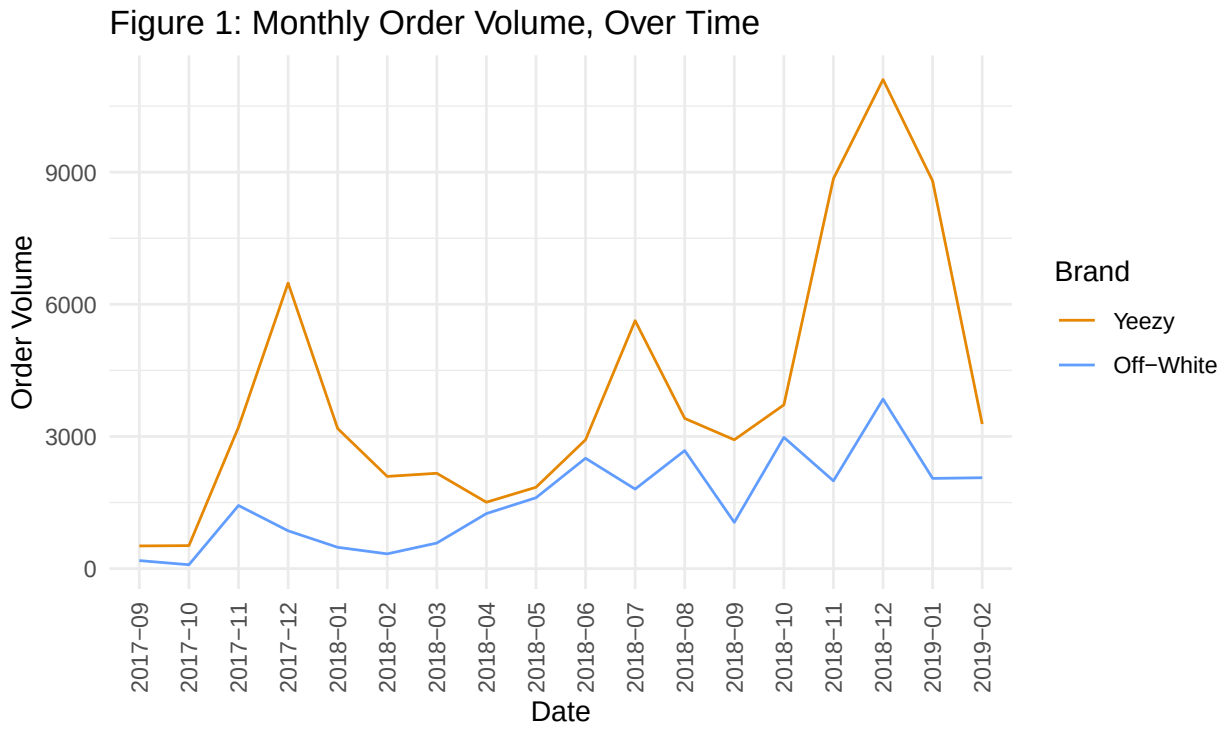
The specific shoe data for this project was collected from the popular resale website StockX. The dataset contains the details of 99,956 orders of *Yeezy* and *Nike X Off-White* shoes made on StockX from September 2017 to February 2019. The variables associated with orders are: *Buyer Region* (State),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

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*,

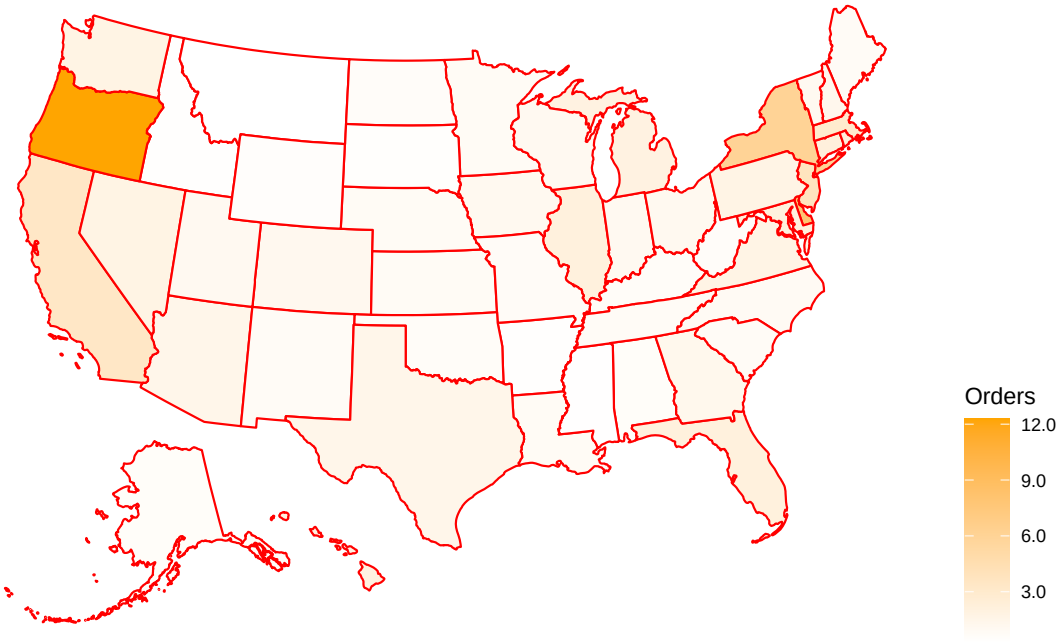
Because certain buying choices could be reflective of economic conditions, we added the variables: *USA Monthly Retail Sales* (Monthly), *State Disposable Income per Capita* (Yearly), and *State Population* (Yearly). These demographic variables were collected from the U.S. Census Bureau, the Federal Reserve, and the Bureau of Economic Analysis.

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.



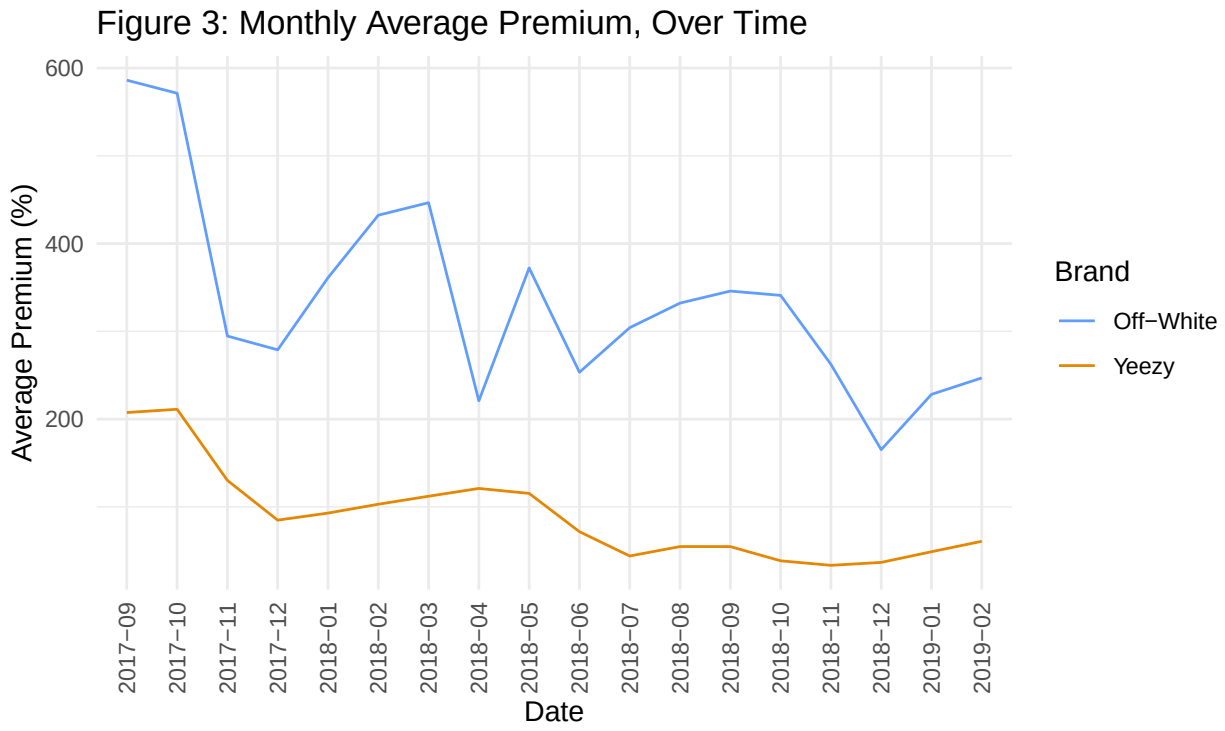
We also wanted to address any geographical component to order volume. After mapping orders, we noticed that Oregon had a disproportional share of orders not explained by population. As shown in *Figure 2*, when controlled for population, Oregon still seems to order the most sneakers out of any state. *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 an average premium of around 284% and *Yeezy* has an average 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.



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*.

Part 2: Model

Because our main goal is to determine the most important factors which impact premium, we decided to employ tree models. We ran both simple decision tree and random forest models. We used the simple decision tree model as a benchmark to determine the effectiveness of Random Forest.

Premium is this case is our dependent variable. With *Brand*, *Sneaker Name*, *Size*, *Buyer Region* (State), *Order Date* (grouped by month), *Primary Color*, *Secondary Color*, and *Material* as our independent variables. The complexity parameter for our decision tree model was placed at .02, minimum observations for split at 300, and max depth at 4. For random forest, our complexity parameter was set at .002 and the number of trees set to 300. For cross-validation, our data was split into testing and training sets, with 20% of the data reserved for testing.

We also decided to attempt to address "Oregon problem" by traditional methods of data manipulation, controlling for population and disposable income and seeing if there was something within the data that could explain Oregon's curious position in order numbers.

Results

Figure 4: Random Forest Variable Importance

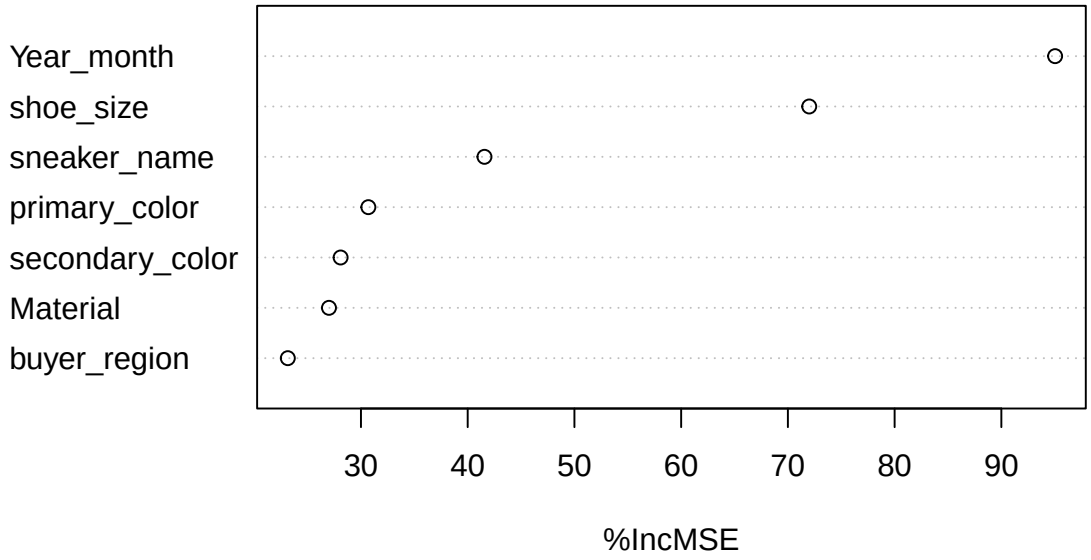


Figure 5: Partial Dependence on Shoe Size

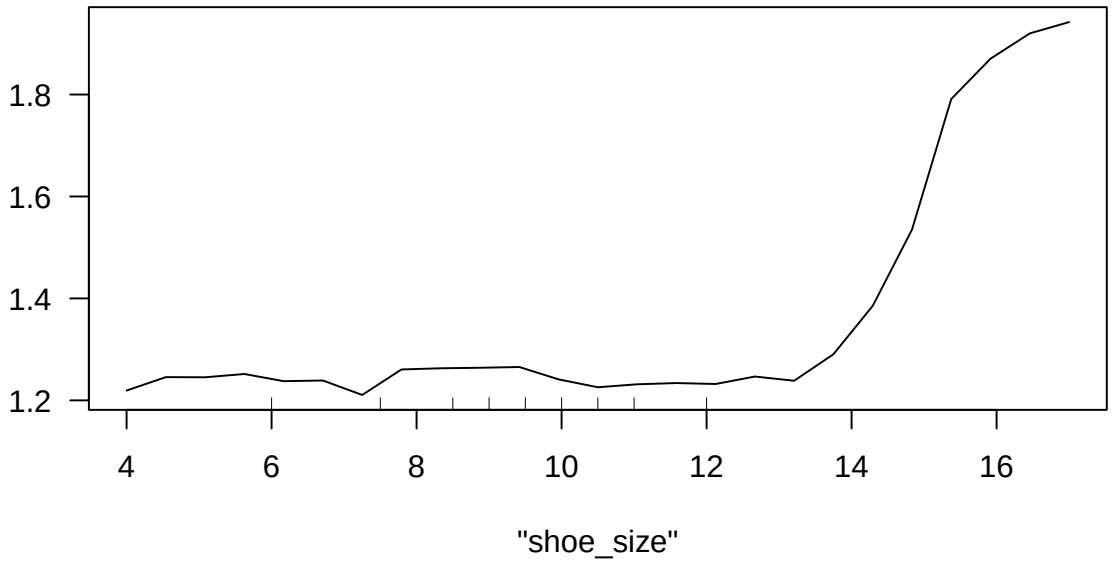
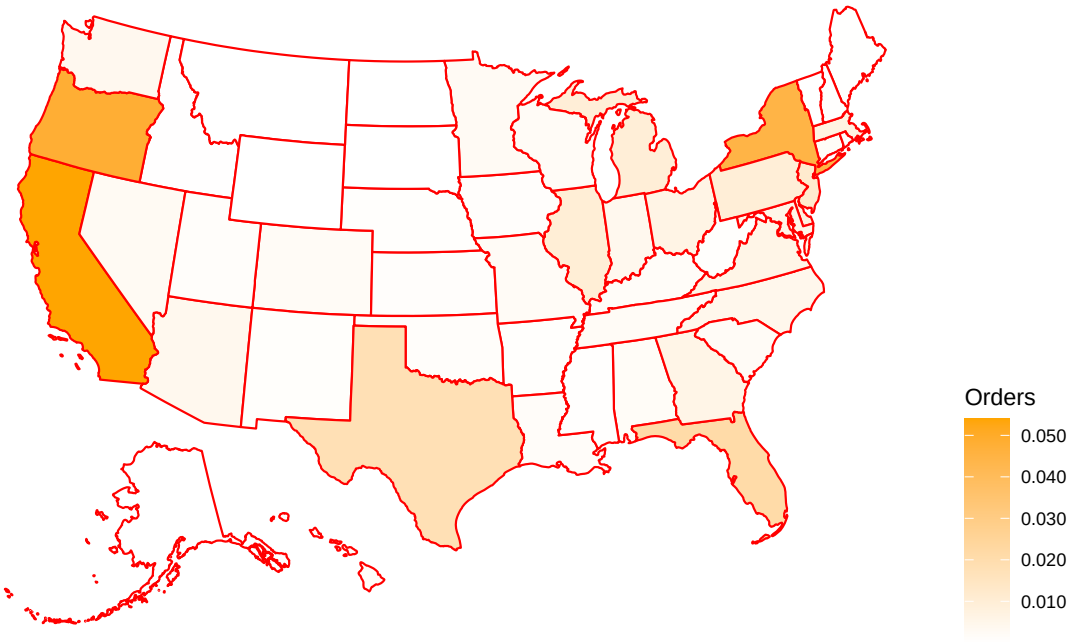


Figure 6: 2019 Adjusted Order Volume by Disposable Income



Conclusion

Looking at *Figure 4*, which is our variable importance plot, our random forest model tells us that the most important variable to determining premium is *Order Date*. This result is not surprising due to the seasonal nature of the retail market. Based on our average premium plot (*Figure 3*), it seems that the best way to maximize premium is to sell during the 'off season' that is: avoid holidays and restocks. The RMSE for our Random Forest model was 31.6% and the RMSE for the decision tree was 65.7%.

The next most important variable in determining premium is *Size*. There could be a number of reasons for this result. Based on *Figure F* in the Appendix, we can see that premium can be afftected by size. Interestingly, there appears to be a distinction between half sizes and full sizes for *Nike X Off-White* shoes, with full sizes between 8-10 commanding higher premiums than half sizes. Sizes over 14 for both *Nike X Off-White* and *Yeezy* command higher premiums, however *Yeezy* doesn't show the same effect as *Nike X Off-White* in sizes below 14. The effect below size 14 could be due to sheer popularity of sizes or due to a supply issue, there is a possibility that *Nike X Off-White* makes more shoes in half sizes than full sizes. The effect over size 14 seems to be cause by both demand and supply, not a lot people have over size 14 feet and brands aren't incentivized to produce many styles of sizes that are so far above average.

Figure 5 appears to confirm what we saw in *Figure F*, which shows a spike in the partial dependence plot between sizes 8-10 and after size 14.

The next most important variable was *Sneaker Name*, this refers to the specific style of shoe. This captures buyers' preferences for one sneaker style over the other. We also noticed that release date was captured using this variable since each style has a different release date. Over the course of the project, we had originally intended to use *Brand* as part of the model, but we quickly noticed that *Sneaker Name* captures *Brand*, because brands have distinct names for their styles.

Partial Dependence plots for *Order Date* and *Sneaker Name* are included in the appendix as *Figure G* and *Figure H*, respectively.

The color and material variables *Primary Color*,*Secondary Color* and *Material* were the next most important, this isolates any effects color and material may have on buyer preferences that isn't captured by the name of the style.

The variable of least importance in the model was *Buyer Region*. This shows that even though there appeared to be customers in some states that purchased more sneakers when adjusted for population or paid a higher premium on average, geographic location did not have have as large of an impact on premium amounts as shoe characteristics.

So what does this tell us? Well, if a seller wants to maximize their premium they should focus on selling *Nike X Off-White* sneakers in sizes 15 and above during lulls in holidays and restocks/releases. Obviously, one cannot corner market on abnormal shoe sizes in the right tail, so a more reasonable strategy would be to focus on selling sneakers in the 8-10 range, focusing on full sizes. *Yeezy* sneakers don't seem to have a similar pattern when it comes to sizing, only really increasing in premium at sizes past 14. This leads us to conclude that desirability of certain styles may be more important in determining premiums for *Yeezy*.

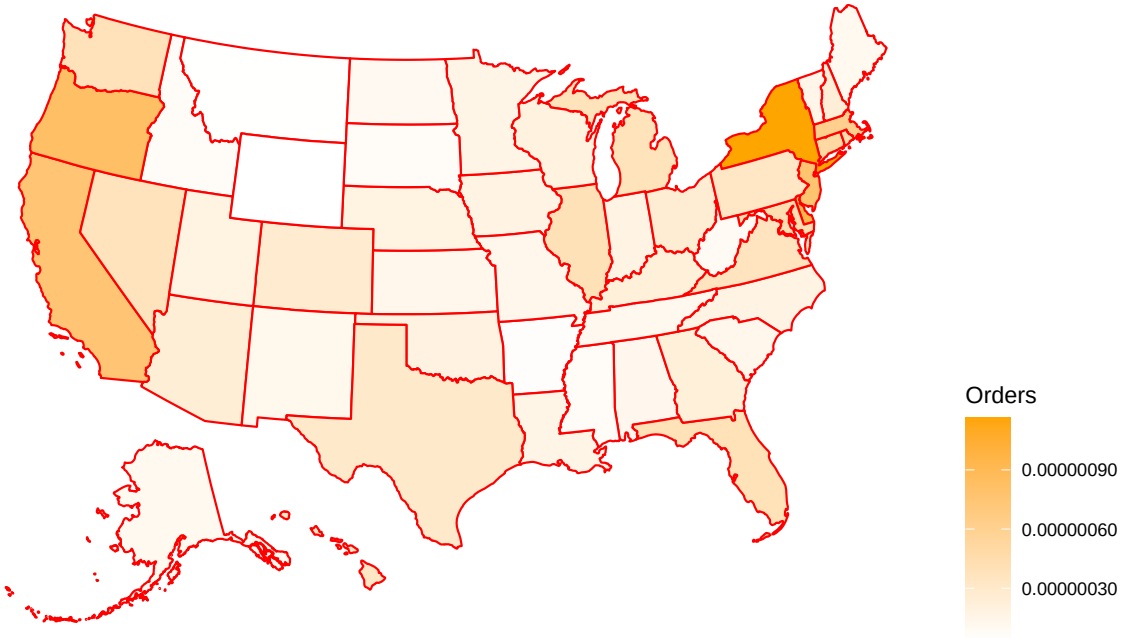
Back on the Oregon Trail

Addressing the "Oregon Issue", we re-grpahed the order numbers and controlled for disposable income per capita of each state. Meaning we were trying to account for any income-related effects as to why Oregon had such a high number of orders. As you can see in *Figure 6*, it appears that even after controlling for income, Oregon still has a disproportionate amount of orders compared to the rest of the 50 states.

What can we conclude from this? Well, there appears to be some sort of noise our data isn't accounting for. After doing a bit of research into Oregon, it seems that Oregon's sneaker culture could possibly explain its large share of orders. The Portland area is also home to Nike and the American headquarters of Adidas.

Appendix

Figure A: 2017 Total Order Count per 10000 Persons



2019 Total Order Count per 10000 Persons

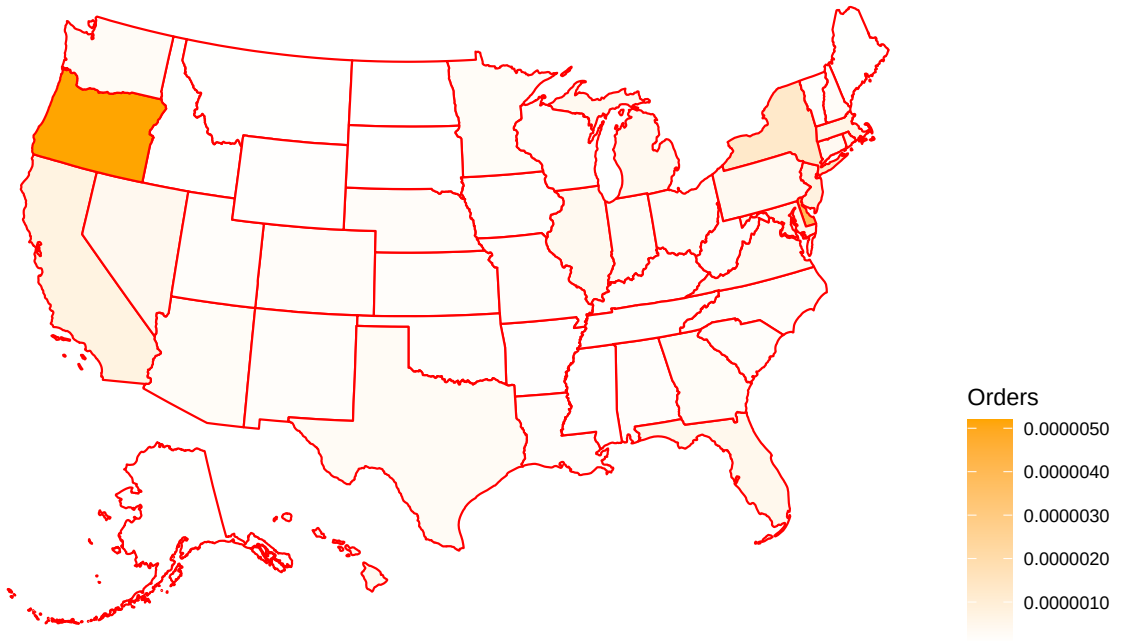


Figure C: 2017 Average Resale Premium by State

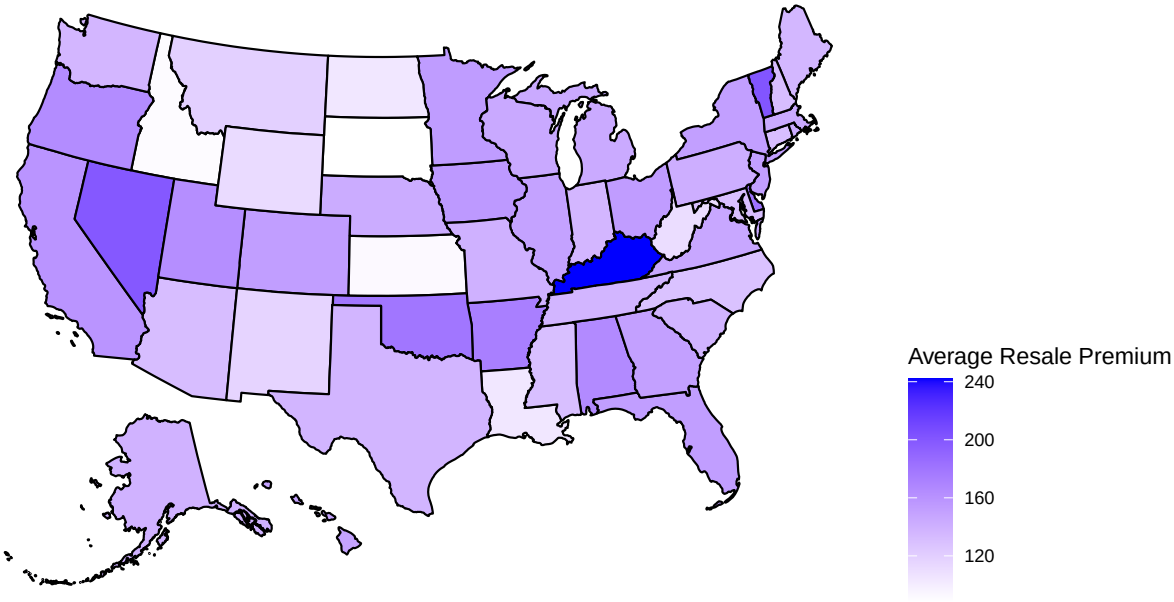


Figure D: 2018 Average Resale Premium by State

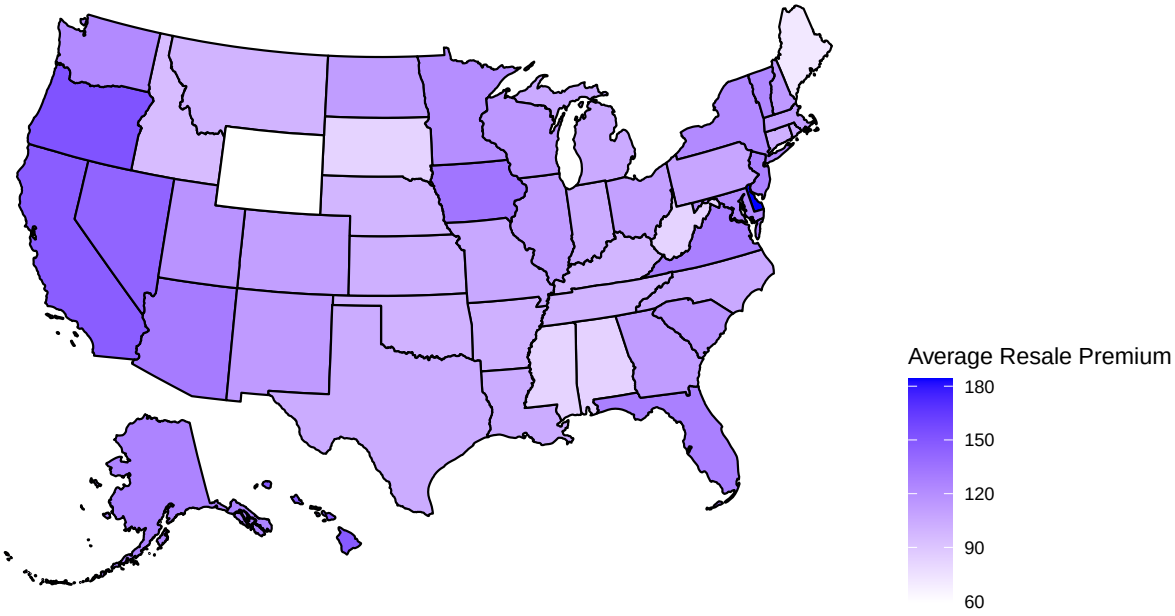


Figure E: 2019 Average Resale Premium by State

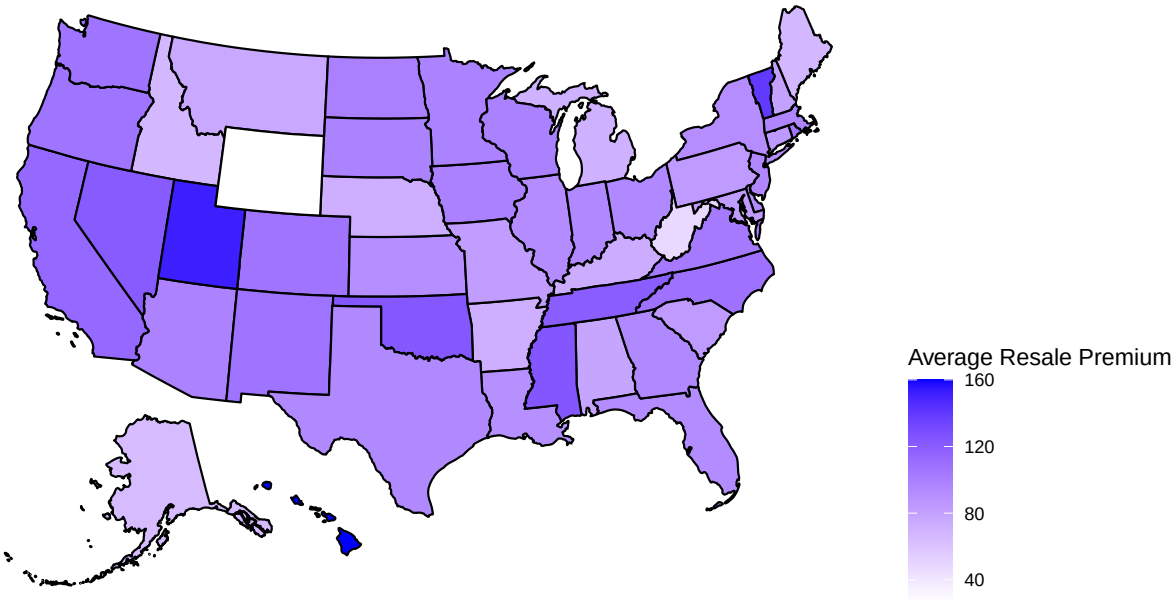


Figure F: Average Premium by Shoe Size

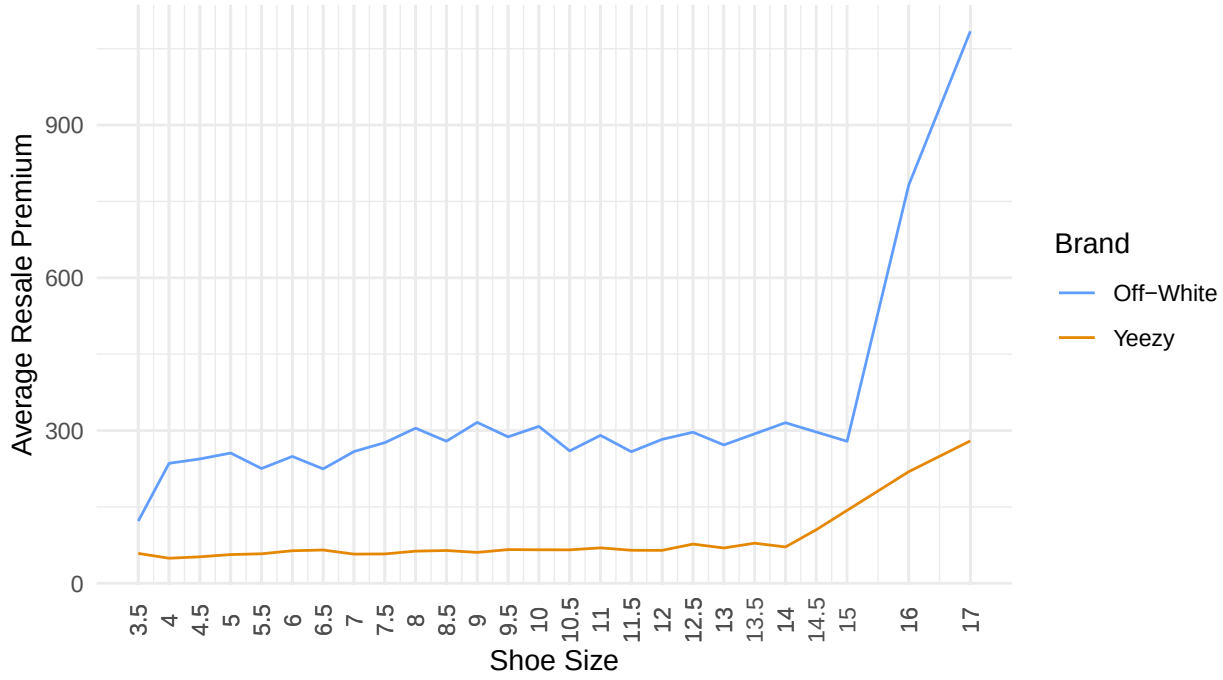


Figure G: Partial Dependence on Year−Month

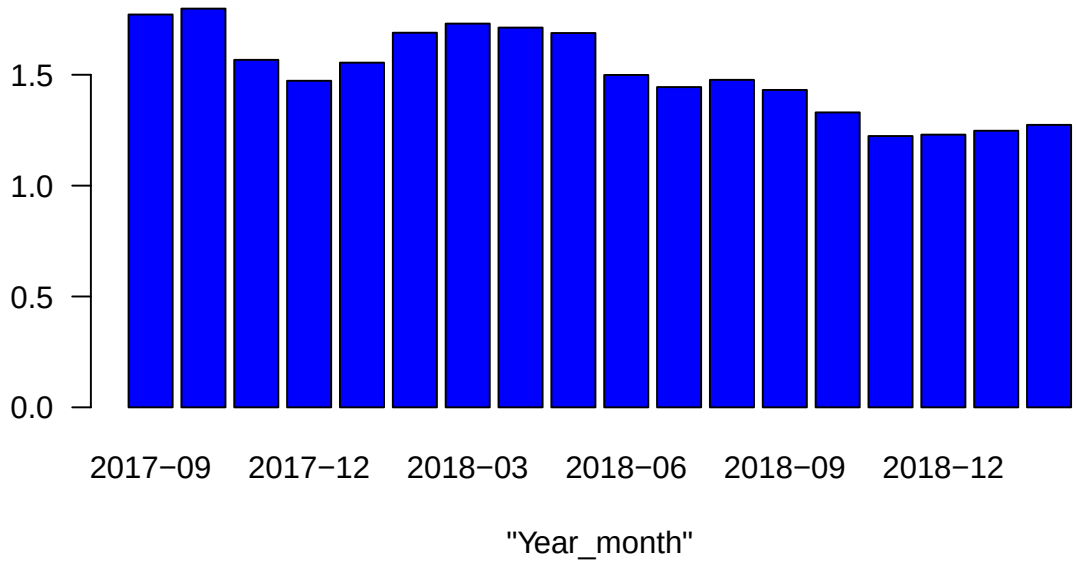


Figure H: Partial Dependence on Sneaker Name